Whole Genome Epidemiological Typing of Escherichia coli

Kaas, Rolf Sommer

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Whole Genome Epidemiological Typing of Escherichia coli

Rolf Sommer Kaas
PhD Thesis
2014

DTU Food
National Food Institute

Supervisors and Funding

This thesis is written in collaboration with three institutions: DTU Food, DTU Center for Biological Sequence analysis (CBS), and Statens Serum Institute. The main supervisor was Frank Møller Aarestrup (DTU Food). Co-supervisor on the first half of the PhD was David W. Ussery (CBS) and co-supervisor on the second half was Ole Lund (CBS). The PhD was supported by the Center for Genomic Epidemiology (www.genomicepidemiology.org) grant 09-067103/DSF from the Danish Council for Strategic Research.

Table of Contents

Acknowledgements

The most important person to thank is of cause my awe-inspiring wife Chilie Maria Sommer Kaas. Chilie gave birth to our lovely daughter half way through this PhD and having a baby and an absent minded, busy husband cannot have been easy. But she has remained supportive and even managed to travel with me on my external research stay to the United States, so that I didn't have to leave my only 1 year old daughter for several months.

I want to thank Frank M. Aarestrup for including me in his wild ambitions. In the beginning I felt like we were running really fast towards a goal we weren't sure existed down a path completely covered in impenetrable fog. However, as time has passed, the path remained, the goal became clearer, and the fog slowly started to lift. I am truly excited to do research with Frank and exciting research requires a leap of faith from time to time.

I am also really thankful that Ole Lund stepped in as co-supervisor after David left for Oakridge National Lab. Ole has many great ideas, and some of them even have to do with science. Ole has been the key in development of several of the bioinformatic methods.

As mentioned, David Ussery left for Oakridge, but he still deserves huge thanks for including me in exciting projects, and not least his dedication to his students.

I also want to thank Rene Hendriksen and Henrik Hasman for including me in several very exciting research projects, which has ended up in several of the publications not included in this PhD.

Whenever administrative tasks seemed confusing or overwhelming Vibeke Hammer stepped in and made everything better. Thanks to Vibeke for relieving me of many administrative headaches.

Thank you also to Carsten Friis who was the first real bioinformatician in Franks group. Carsten took very good care of me when I started here, and made sure to introduce me to all the right people.

A huge thanks also goes to Mette Christiansen and Maria Seier-Petersen who also received me with open arms and made me feel right at home. Mette and Maria showed me that there was more to a PhD than studying and writing.

I would also like to thank Marlene Hansen, who always claims that she can't contribute anything to my PhD, but nonetheless has. Marlene provided me with great articles to read but also gives great feedback both professionally and personally, which is an important skill when you share an office. This leads me to Pimlapas "Shinny" Leekitcharoenphon who I also owe many thanks, for all her help with several articles and always spreading some good mood in the office.

As a bioinformatician I am most dependent on the people I see the least during my workday – the technicians. Many thanks go to all the technicians who made sure that there were actually sequence data for me to work with.

Thanks to Katrine Joensen and Ea Zankari for being my surrogate office colleagues when Shinny and Marlene were absent.

Finally thanks to all of department G for providing a great working environment.

List of original articles

I. **Kaas RS**, Friis C, Ussery DW, Aarestrup FM (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genomics 13:577.

II. Leekitcharoenphon P, **Kaas RS**, Thomsen MCF, Friis C, Rasmussen S, Aarestrup FM (2012) snpTree - a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics 13(Suppl 7):S6.

III. **Kaas RS**, Leekitcharoenphon P, Aarestrup FM, Lund O (2014) Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. PLoS ONE 9(8): e104984.

IV. **Kaas RS**, Rasmussen S, Scheutz F, Lund O, Aarestrup FM (2014) Investigation of methods to define *Escherichia coli* outbreak strains based on whole genome sequence data from 10 different outbreaks. *Manuscript for submission to: J Clin Microbiol.*

List of original articles not included in PhD

Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica. Leekitcharoenphon, Pimlapas; Nielsen, Eva M.; **Kaas, Rolf Sommer**; Lund, Ole; Aarestrup, Frank Møller. In: PLoS One, Vol. 9, No. 2, e87991, 2014.

Genome-Wide High-Throughput Screening to Investigate Essential Genes Involved in Methicillin-Resistant Staphylococcus aureus Sequence Type 398 Survival. Christiansen, Mette Theilgaard; **Kaas, Rolf Sommer**; Chaudhuri, Roy R.; Holmes, Mark A.; Hasman, Henrik; Aarestrup, Frank Møller. In: PLoS One, Vol. 9, No. 2, e89018, 2014.

Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nielsen, Henrik Bjørn; Almeida, Mathieu; Juncker, Agnieszka; Rasmussen, Simon; Li, Junhua; Sunagawa, Shinichi; Plichta, Damian Rafal; Gautier, Laurent; Pedersen, Anders Gorm; Emmanuelle, Le Chatelier; Pelletier, Eric; Bonde, Ida; Nielsen, Trine; Manichanh, Chaysavanh; Arumugam, Manimozhiyan; Batto, Jean-Michel; dos Santos, Marcelo Bertalan Quintanilha; Blom, Nikolaj; Borruel, Natalia; Burgdorf, Kristoffer S.; Boumezbeur, Fouad; Casellas, Francesc; Doré, Joël; Dworzynski, Piotr; Guarner, Francisco; Hansen, Torben; Hildebrand, Falk; **Kaas, Rolf Sommer**; Kennedy, Sean; Kristiansen, Karsten; Kultima, Jens Roat; Léonard, Pierre; Levenez, Florence; Lund, Ole; Moumen, Bouziane; Denis, Le Paslier; Pons, Nicolas; Pedersen, Oluf; Prifti, Edi; Qin, Junjie; Raes, Jeroen; Sørensen, Søren; Tap, Julien; Tims, Sebastian; Ussery, David; Yamada, Takuji; Renault, Pierre; Sicheritz-Pontén, Thomas; Bork, Peer; Wang, Jun; Brunak, Søren; Ehrlich, S. Dusko. In: Nature Biotechnology, 2014.

Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. Joensen, Katrine Grimstrup; Scheutz, Flemming; Lund, Ole; Hasman, Henrik; **Kaas, Rolf Sommer**; Nielsen, Eva M.; Aarestrup, Frank Møller. In: Journal of Clinical Microbiology, Vol. 52, No. 5, 2014, p. 1501-1510.

Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. Zankari, Ea; Hasman, Henrik; **Kaas, Rolf Sommer**; Seyfarth, Anne Mette; Agersø, Yvonne; Lund, Ole; Larsen, Mette Voldby; Aarestrup, Frank Møller. In: Journal of Antimicrobial Chemotherapy, Vol. 68, No. 4, 2013, p. 771-777.

Veillonella, Firmicutes: Microbes disguised as Gram negatives. Vesth, Tammi Camilla; Ozen, Asli; Andersen, Sandra Christine; **Kaas, Rolf Sommer**; Lukjancenko, Oksana; Bohlin, Jon; Nookaew, Intawat; Wassenaar, Trudy M.; Ussery, David. In: Standards in Genomic Sciences, Vol. 9, No. 2, 2013, p. 431-448.

Draft Genome Sequence of the Yeast Pachysolen tannophilus CBS 4044/NRRL Y-2460. Liu, Xiaoying; **Kaas, Rolf Sommer**; Jensen, Peter Ruhdal; Workman, Mhairi. In: Eukaryotic Cell (Online Edition), Vol. 11, No. 6, 2012, p. 827.

European freshwater VHSV genotype Ia isolates divide into two distinct subpopulations. Kahns, Søren; Skall, Helle Frank; **Kaas, Rolf Sommer**; Korsholm, H.; Jensen, Ann Britt Bang; Jonstrup, Søren Peter; Dodge, M. J.; Einer-Jensen, Katja; Stone, D.; Olesen, Niels Jørgen. In: Diseases of Aquatic Organisms, Vol. 99, No. 1, 2012, p. 23-35.

Population Genetics of Vibrio cholerae from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. Hendriksen, Rene S.; Price, Lance B.; Schupp, James M.; Gillece, John D.; **Kaas, Rolf Sommer**; Engelthaler, David M.; Bortolaia, Valeria; Pearson, Talima; Waters, Andrew E.; Prasad Upadhyay, Bishnu; Devi Shrestha, Sirjana; Adhikari, Shailaja; Shakya, Geeta; Keim, Paul S.; Aarestrup, Frank Møller. In: mBio, Vol. 2, No. 4, 2011, p. e00157.

Summary

Escherichia coli (*E. coli*) is of huge importance in global health both as a commensal organism living within its host or as a pathogen causing millions of infections each year. Infections occur both sporadic and as outbreaks with sometimes up to thousands of infected people. To limit the number of infections it is important to monitor pathogenic *E. coli* in order to detect outbreaks as quickly as possible and find the source of the outbreak. The effectiveness of monitoring and tracking of pathogens is very dependent on the typing methods that are employed. Classical typing methods employed for *E. coli* is in general expensive and to some extent unreliable. Next generation sequencing has quickly become a tool widely available and has enabled even smaller laboratories to do whole genome sequencing (WGS). Having the entire genome available provides the opportunity to create the ultimate typing method. This PhD thesis attempts to take the first steps toward such a method.

In **Kaas I** all publicly available *E. coli* genomes sequenced (186) are analyzed. 1,702 core genes were found in all genomes. 3,051 genes were found in 95% of the genomes. The pan genome was found to consist of 16,373 genes. The overall phylogeny was inferred from the core genome and also set into context of the *Escherichia* genus. The variance within each gene cluster was calculated in order to compare the variance between genes and possibly identify typing targets for further study. The variance scores calculated was also used to compare the three MLST schemes that exist for *E. coli*.

It quickly became clear that single nucleotide polymorphism (SNP) analysis was becoming the method of choice for inferring the phylogeny of bacterial outbreaks. However, the method remained unavailable to many people due to technical obstacles. In **Kaas II** we describe the SNP method and the validation behind a web

server that we set up in order to overcome some of the technical obstacles faced by many people and thereby making the method more available. The method briefly, calls SNPs against a specified reference sequence, creates an alignment (pseudo-sequence) of all the SNPs, and uses the maximum likelihood (ML) method to create a tree. The most important detail in the method is the assumption made about "missing" SNPs. Meaning SNPs called in one strain but not in another. It was assumed that SNPs not found in a position was due to that nucleotide being identical to the one in the reference sequence. The assumption is in general valid if all the genomes compared are closely related and the sequencing data is of good quality.

In **Kaas III** we sought to overcome the assumption mentioned above but most important of all we wanted to create a method that could handle sequence data obtained from different sequencing technologies. The method from **Kaas II** was completely rewritten and a new web server (CSI Phylogeny) was published that could handle sequence data of all kinds and no longer made assumptions about missing SNPs. Very briefly, the method differs from **Kaas II** mainly by validating all the locations in all the genomes in which a SNP has been called in any genome. In parallel to the development of a new SNP method another method was also developed that briefly, relies on counting nucleotide differences (ND) between each genome pair, while also validating each position analyzed and ignoring the positions that cannot be validated thereby creating a distance matrix that is used as input to an UPGMA method that creates the final phylogeny. The ND method was also implemented as a web server and published.

If whole genome sequencing is to be used for routine monitoring and tracking of *E. coli* pathogens, it is crucial to have an idea of how large the difference is between isolates from the same outbreak, compared to the difference to other non-outbreak

isolates, in order to do reliable distinctions. In **Kaas IV** we analyzed ten different outbreaks. Seven of the outbreaks were sequenced for the study and three of the outbreaks were obtained from published studies. Several background isolates that resembled the outbreak isolates were also sequenced. Five different bioinformatic methods were evaluated against the 10 outbreaks. The five different methods were based on SNP, ND, core genes, k-mers, and average nucleotide identity (ANI). Only the ANI method was not able to cluster all outbreaks correctly. The pairwise distance between all isolates were also calculated by each method and compared. Most methods showed lower distance between isolates in the same outbreak compared to the background strains, but only the SNP method was able to set one common threshold for outbreak isolates versus non-outbreak isolates for the entire dataset.

Whole genome sequencing is a powerful but also a rather new tool. This PhD thesis has hopefully shed some light on how we can continue development of whole genome sequence typing and also made WGS more available to a broader audience.

Danish Summary

Escherichia coli (*E. coli*) spiller en vigtig rolle i den globale sundhed både grundet dennes rolle som kommensal bakterie, der lever i dennes vært og som patogen bakterie, der er skyld i millioner af infektioner hvert eneste år. Infektionerne er både sporadiske eller som udbrud med tusindvis af smittede i visse tilfælde. For at mindske antallet af infektioner er det vigtigt at overvåge patogene *E. coli* med henblik på hurtigt opdagelse af udbrud og sporing af kilden til disse. Effektiviteten af overvågning og sporring er i høj grad afhængig af typningsmetoderne der anvendes. De klassiske typningsmetoder, der anvendes til *E. coli* er overordnet set dyre og til en hvis grad ikke helt til at stole på. Næste generations sekventering er hurtigt blevet vidt tilgængelig og har gjort det muligt for selv mindre laboratorier at udnytte hel genom sekventering. At have hele bakteriegenomet tilgængeligt giver nu mulighed for at udvikle den ultimative typningsmetode. Denne Ph.d. afhandling forsøger at tage de første skridt imod en sådan metode.

I **Kaas I** analyseres alle offentligt tilgængelige sekventerede *E. coli* genomer (186). 1.702 kernegener blev fundet i alle genomer. 3.051 gener blev fundet i 95% af alle genomer. Pan-genomet blev beregnet til at indeholde 16.373 gener. Den overordnede fylogeni blev estimeret fra kernegenerne og også sat i kontekst til genusset Escherichia. Variansen i hvert gen blev beregnet med henblik på at sammenligne variansen i mellem forskellige gener og identificere mulige typningsmarkører til yderligere undersøgelse. De beregnede variansscorer blev også brugt til at sammenligne de tre MLST skemaer, der eksisterer for *E. coli*.

Det stod hurtigt klart at Single Nukleotid Polymorfisme (SNP) var ved at blive den fortrukne metode til at udlede et bakterieudbruds fylogeni. Dog forblev metoden utilgængelig for mange, grundet tekniske forhindringer. Vi beskriver i **Kaas II** en

SNP metode og en validering af denne for en webserver vi implementerede netop for at overkomme nogle af disse tekniske forhindringer og dermed øge tilgængelighed af denne metode. Metoden i korte træk: Der kaldes SNPs imod en referencesekvens, disse SNPs bliver så sat sammen til et ”alignment” (pseudo-sekvens) og ved brug af metoden ”Maximum Likelihood” udledes et fylogenetisk træ. Den vigtigste detalje i denne metode er antagelsen der laves omkring ”manglende” SNPs. Med dette menes SNPs, der er kaldt i en stamme, men ikke i en anden. Det blev antaget at grunden til at en SNP manglede i en position var at nukleotiden i denne position var identisk med nukleotiden i samme position i referencesekvensen. Antagelsen holder så længe at stammerne der bliver sammenlignet er meget ens og sekvensdata er af god kvalitet.

I **Kaas III** søgte vi at overkomme førnævnte antagelse, men vigtigst af alt ville vi gerne have en metoder der kunne håndtere sekvensdata opnået via forskellige sekvensteknologier. Metoden fra **Kaas II** blev fuldstændig omskrevet og en ny web server (CSI Phylogeny) blev publiceret, der kunne håndtere alle former for sekvensdata og ikke længere foretog nogen antagelser vedrørende manglende SNPs. Meget kort, så adskiller metoden hovedsageligt sig fra **Kaas II** ved at validere alle positioner i alle genomer, hvor SNPs er blevet kaldt. Parallelt med udviklingen af den nye SNP metode, blev også udviklet en metode, der kort fortalt, bygger på at tælle antallet af nukleotid forskelle (ND) mellem hvert genom par, samtidig med at der også laves positions validering og positioner der ikke kan valideres ignoreres. Derved skabes en distance matrix, der bliver brugt som input til UPGMA metoden der udleder den endelige fylogeni. ND metoden blev også implementeret som web server og publiceret.

Hvis hel genom sekventering skal bruges til rutine overvågning og sporing af patogene *E. coli*, så er det afgørende at vide hvor stor forskel der kan forventes at

findes blandt isolater i samme udbrud i forhold til isolater, der ikke er en del af et udbrud, hvis man skal være i stand til at kunne skelne. Vi analyserede 10 forskellige udbrud i **Kaas IV**. Syv af disse udbrud blev sekventeret til dette studie og tre af udbruddene blev hentet fra publicerede studier. Flere baggrundsisolater, der ligner udbrudsisolaterne blev også sekventeret. Fem forskellige bioinformatiske metoder blev evalueret på de ti udbrud. De fem forskellige metoder var baseret på SNP, ND, kernegener, k-mers og gennemsnitlig nukleotid identitet (Average Nucleotide Identity – ANI). Kun metoden ANI, kunne ikke klynge alle isolaterne i deres respektive udbrud. Den parvise afstand imellem alle isolater blev også beregnet med hver enkel metode og sammenlignet. De fleste metoder beregnede lavere afstand imellem isolater i samme udbrud end til baggrundsisolater, dog var det kun SNP metoden, der var i stand til at sætte én fælles tærskel for udbrudsisolater versus ikke-udbrudsisolater for hele datasættet.

Hel genom sekventering er et kraftfuldt, men også en ret nyt værktøj. Denne Ph.d. afhandling har forhåbentlig været med til at kaste lys over hvordan vi kan fortsætte udviklingen af hel genom typning og gjort hel genom sekventering mere tilgængeligt for et bredere publikum.

Problem Statement

The first next generation sequencer was released around 2004 and triggered a massive increase in whole genome sequencing (WGS), providing new and powerful ways to obtain insights into genomics. While sequencing technology has evolved at an impressive speed, becoming even faster and cheaper, WGS is still a young field with lots of questions to be answered. In this PhD we attempt to elucidate several issues regarding the use of WGS as a tool for typing of *E. coli*.

E. coli was chosen due to its role as a frequent human, animal, and food borne pathogen and its genomic profile that is clonal but still very diverse. The PhD aimed at exploring how diverse the genome of the available whole genome sequences were and to estimate the variance of the genes. The objective of the variance analysis was to find potential epidemiological markers for future studies.

Single Nucleotide Polymorphisms (SNPs) has been widely used to describe different bacterial outbreaks and recent evolution. It was the aim of this thesis to analyze results obtained by the SNP method, but also alternative methods in order to evaluate their use for typing purposes. In general, there is a need in WGS to establish standards for the definition of bacterial clones.

The project is part of the Center for Genomic Epidemiology (CGE). The main objective of CGE is to facilitate global surveillance of microbial pathogens using WGS. From this objective also follows the need to make WGS tools available to clinical laboratories and epidemiologists in order to make WGS a feasible alternative to classical methods. An additional objective of this thesis was therefore also to make user-friendly tools available for complete SNP analysis and also to improve on the existing method to make it platform (sequence) independent.

E. coli

Taxonomy

In 1885 Theodor Escherichia discovered a new species in the feces of healthy individuals. It was originally named *Bacterium coli commune*, later reclassified *Bacillus coli* before it was finally classified *Escherichia coli* (*E. coli*) [1].

E. coli belongs to the family *Enterobacteriaceae* and are non-spore forming, facultative anaerobe, and Gram-negative rods. *E. coli* can be both motile and non-motile. Motile *E. coli* has peritrichous flagella [2].

Since Nobel price winners Edward Tatum and Joshua Lederberg used the *E. coli* K12 strain, as a model organism to show bacterial conjugation in the late 1940s it has been the preferred model organism in Biology research. Researchers in bacterial genetics, biochemistry, and physiology have come to favor *E. coli* K12 due to its accessibility, rapid and simple laboratory growth conditions, low virulence, tractable genetics, and metabolic versatility [3].

Ecology

E. coli exists both as commensal and as a widespread pathogen. Commensal *E. coli* are predominantly found in the gut of mammals, but in general exists in warm-blooded animals and reptiles [4,5]. However, *E. coli* has the ability to survive in the environment (e.g. water) for prolonged periods of time and is often used as an indicator for fecal contamination. In recent years more focus has been given to the environmental *E. coli* and it has been estimated that half of the *E. coli* population resides in secondary habitats. Studies have also shown that specific strains of *E. coli* (those capable of saprophytism) are capable of growing under these environmental conditions [5]. Walk *et al.* reported 5 new phylogenetic clades [6], from strains

isolated mainly from the environment that was distinct from known *E. coli* but could not be distinguished using traditional phenotypic profiling [6,7] [**Kaas I**]. The primary habitat of commensal *E. coli* is in the large intestine of the digestive tract and predominantly in the caecum and the colon. They are populating the mucus layer that covers the epithelial cells throughout the digestive tract. The mucus gets degraded and shed in to the intestinal lumen where it gets excreted in the feces. The competition in the microbiota in the gut is high and *E. coli is* outnumbered 100/1 to 10,000/1 [5]. *E. coli* has adapted its metabolism to utilize sugars from the mucus; especially gluconate seems to play an important role. Commensal *E. coli* benefits from its relationship with the host owing to a steady flow of nutrients, a stable environment, protection from some stress factors, and also gains transport and dissemination. The host benefits are more implicit. It is understood that *E. coli* can provide colonization resistance, but *E. coli* does also play a part in the eco system of the microbiota, primarily by limiting oxygen in the environment, and thereby enhancing the conditions for anaerobe organisms, in turn these anaerobe organism might further benefit the host. However, this is a discussion that is well beyond the scope of this thesis [5].

Pathogenic Classification

In immunocompromised hosts, where the gastrointestinal barrier is breached commensal *E. coli* can cause disease (ex. peritonitis). However, the vast majority of *E. coli* infections are due to pathogenic *E. coli* that has adapted a broad range of virulence factors, which leads to a wide spectrum of diseases (See Figure 1). Pathogenic *E. coli* can be divided into two main categories: diarrhoeagenic *E. coli* and extraintestinal *E. coli* (ExPEC). The majority of diarrhoeagenic *E. coli* can further be divided into six pathovars: enteropathogenic E. coli (EPEC), enterohaemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC),

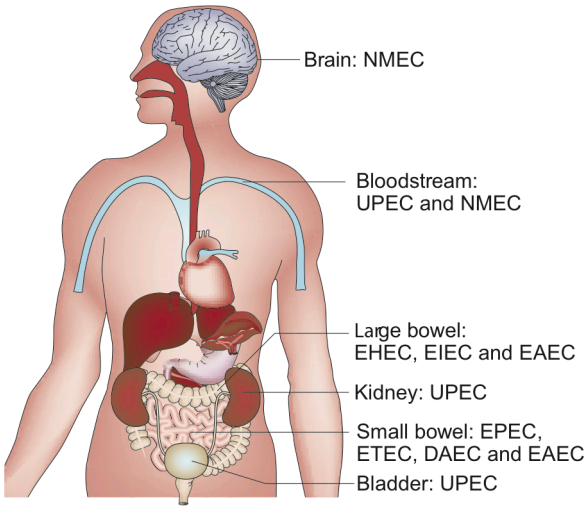

Figure 1. Colonization sites of pathogenic *E. coli*. Figure is borrowed from Croxen et al. [9].

enteroinvasive *E. coli* (EIEC), and diffusely adherent *E. coli* (DAEC). It is widely accepted that *Shigella spp.* falls within the species *E. coli* [8] and it should be classified in the pathovar EIEC [9]. However, due to the clinical significance of *Shigella* the nomenclature is still maintained [10]. The two most common ExPEC pathovars are uropathogenic *E. coli* (UPEC) and neonatal meningitis *E. coli* (NMEC), others exists but will not be described here. It is important to keep in mind that the pathovars describe the pathogenicity of an *E. coli* strain but contains little information on its phylogenetic relationship to other *E. coli* strains. The reason being that a pathovar is largely defined by virulence factors that are often located on mobile genetic elements and are therefore subject to horizontal gene transfer (HGT) [9,10]. A study by Ogura et al. showed how different *E. coli* clones has independently evolved into EHECs [11].

Epidemiology & Clinical importance

E. coli is primarily spread through contaminated water and food. Diarrhoeagenic *E. coli* is a massive problem in developing countries where they are endemic and a significant contributor to childhood mortality. The most important pathovars in developing countries are ETEC, EPEC and EIEC. A rather large infectious dose is required for both ETEC and EPEC infections, which explains why there is practically no direct person-to-person transmission. ETEC is also known as traveller's diarrhea because ETEC is among the most common causes of diarrhea in visitors to

developing countries. Studies have shown that in areas where ETEC is endemic, there also exists thorough contamination, especially in the warm and wet months where the bacteria thrive. It is believed that mucosal immunity can be obtained in exposed individuals and it has been shown that asymptomatic individuals may shed large portions of virulent ETEC in their stools. These two factors explain how ETEC can remain endemic and why it almost only affects visitors and children. While ETEC in developed countries are less predominant among the human population it is a problem among swineherds, in food production, where it causes diarrhea and edema disease. Edema disease is affecting post-weaning pigs, where it is often fatal.

EPEC is like ETEC also primarily found in children. EPEC is found in children under the age of 2 years and primarily in infants younger than 6 months. EPEC does not affect visitors like ETEC does. It is believed that humans loose specific receptors with age and is therefore not affected by EPEC. The primary reservoir is believed to be asymptomatic human adults and children. The vast majority of EPEC infections are found in children but outbreaks of EPEC affecting adults is seen and often the source is contaminated food or water.

EHEC is in contrast to ETEC and EPEC predominantly found in developed countries. EHEC requires low infectious doses and is in addition to spread by contaminated food and water also spread by person-to-person transmission. The major outbreaks caused by EHEC have gained the most attention, but the sporadic infections cause the largest disease burden, as estimated by the Center for Disease Control (CDC) in the US. However, it is probable that insufficient typing fails to reveal a number of these sporadic cases as outbreaks. Stx producing *E. coli* is found in a wide variety of animals that is usually asymptomatic due to lack of receptors that bind shiga toxins. Strains responsible for human disease is manly found in the gut flora of cattle. The

low number of organisms (100-200) required for infection makes cross contamination of food products and seeds a significant problem and has in many incidents been reported to be the source of outbreaks, including the German outbreak in 2011 that was believed to originate from contaminated sprout seeds [12]. The most common EHEC serotype reported is O157:H7 but other serotypes is also important causes of EHEC infections.

EAEC is a very diverse group and contains considerable genetic heterogeneity. This fact along with difficulties in establishing consistent defining factors for this group continues to make the pathogenicity and clinical relevance controversial topics. The main reservoir is believed to be human. EAEC is often isolated from children suffering from diarrhea in developing countries. However EAEC is also often found to co-exist with other pathogenic *E. coli* making its clinical relevance difficult to estimate. EAEC is frequently mentioned in relation to persistent diarrhea, however no solid evidence of this exists [13]. EAEC has also been found to cause urinary tract infections (UTI) in Denmark [14]. A hybrid of an EAEC and an EHEC was believed to cause a large German outbreak in 2011.

Non-Shigella EIEC is less widespread than previously mentioned pathovars although it has been reported to be predominant in some areas [15]. Shigella has received a lot more attention as it continues to cause a very high amount of infections. Kotloff et al. estimated more than 1 million deaths caused by Shigella spp. in 1999 in the developing world [16] and CDC's FoodNet has registered Shigella to be the second most prevalent foodborne pathogen in 2013 with Salmonella being the most prevalent and Campylobacter to be the third most prevalent pathogen [17]. The main reservoir is believed to be human and the very low dose required for infection makes person-to-person transmission and cross contamination important factors. Shigella is found

worldwide although the highest disease burden is found in the developing world, where it is a significant cause of possible fatal diarrhea in especially young children.

The mode of acquisition is largely unknown for DAEC. Some studies suggest that DAEC mostly affect children between the age of 1 and 5 years. The lack of solid pathogenic markers makes the clinical importance of DAEC uncertain [18].

Urinary tract infections (UTIs) are one of the most common community acquired infections and also the most common nosocomial infection. The risk of getting a nosocomial UTI is estimated to be 5-10% each day for patients with catheters but is also very common among non-catheter patients [19]. The predominant cause of UTIs is UPEC *E. coli*, which is estimated to be the cause of 70-90% of the community acquired UTIs and 50% of the nosocomial UTIs [20]. Even though a large portion of UTIs is asymptomatic and UTIs rarely has fatal outcomes, the vast amount of infections makes it a very important pathogen both from a health perspective and due to the significant economic burden it imposes on society. However, actual UPEC outbreaks are rare and most infections are sporadic [13].

NMEC is the second cause of neonatal meningitis. The average onset is about 6-9 days from birth. It is believed that the acquisition of NMEC comes from the gut flora of the mother or the environment. Few risk factors has been identified but around one third of all infections happens in premature infants [21].

Pathogenesis

In general the *E. coli* pathovars affects the same host mechanisms, but do so using very different approaches. EPEC and EHEC both belong to a family of pathogens that form attaching and effacing (A/E) lesions on epithelial cells. The bacteria efface the microvilli and form distinct pedestals on the host cell beneath the attached bacterial cell. The initial attachment of the two pathovars is through different adherence

factors, that are not all fully understood but involves different types of pili and flagella. The intimate attachment happens through the bacterial outer-membrane protein intimin. The intimate attachment recruits other proteins that in turn lead to actin replacement and pedestal formation. The literature describes this process in detail for the prototypes of EHEC and EPEC, but it turns out the two prototype strains studied are not representative for all EHECs and EPECs [22]. Following intimate attachment EPEC pathovars translocate a variety of effector proteins in to the target host cell using the type 3 secretion system (T3SS). The specific effectors are different between the EPEC strains but are known to be responsible for disruption of mitochondrial structure and function, inhibition of phagocytosis, disruption of the tight junction between epithelial cells, reduction of protein trafficking and increased ion secretion. Several of these factors are believed to be the cause of diarrhea also including increased intestinal permeability, intestinal inflammation, active ion secretion and the loss of epithelial surface due to effacement.

EHEC causes both bloody/non-bloody diarrhea and hemolytic uremic syndrome (HUS) in humans. The Stx toxins also known as verocytotoxins (VTs) are the primary virulence factors of EHEC pathovars. *E. coli* strains that produce Stx/VT are also known as Shiga toxin producing *E. coli* (STEC) and verocytotoxin producing *E. coli* (VTEC), respectively. Two types of Stx toxins are relevant for human infections; Stx1 and Stx2. EHECs lack a secretory system for Stx and are only released through phage-mediated lysis, which is why antibiotic therapy should be discouraged in relation to an EHEC infection. The receptors of Stx are found on Paneth cells in the human intestinal tract. The uptake and subsequent activation of the toxin in these cells are believed to prevent protein synthesis and lead to necrosis and cell death. From the epithelial cells Stx enters the bloodstream where it is transported to the kidneys,

where it can lead to HUS that in turn can lead to fatal acute renal failure. Stx is also found in *Shigella dysenteriae* that is also able to cause HUS, but not other species of *Shigella*. More than 200 serotypes of *E. coli* produce Stx toxins (VTEC/STEC). However, many lack the pathogenicity island known as the locus of enterocyte effacement (LEE) and are therefore not causing disease in humans. Only the VTEC/STEC, containing this island, lead to human disease and is categorized EHEC.

Some members of the ETEC pathovar also produce Stx toxins and are known as ETEC/STEC and are an important pathogen of pigs, where it colonizes the small intestine. The toxins enter the bloodstream and bind specific receptors on epithelial and endothelial cells. The toxin impairs blood vessels and leads to edema, ataxia, and death [23].

The pathovar ETEC causes watery diarrhea in humans. ETEC enterotoxins consist of two groups: heat-labile enterotoxins (LT) and heat-stabile enterotoxins (ST). ETEC strains express either one of the toxins or both. LT toxins increase intracellular cAMP that leads to increased Cl^- secretion from the epithelial cell, which leads to diarrhea. ST toxins exists in two unrelated classes: STa and STb. Only STa causes disease in humans, STb causes disease in animals. STa causes increased levels of cGMP in the host cell that leads to increased secretion from the cell and causes diarrhea. STb causes elevation of Ca^{2+} concentration in the host cell that leads to increased ion secretion. Interestingly, it has been suggested that there is a link between countries with a high prevalence of ETEC and a low rate of colon cancer [24].

The knowledge of EAEC pathogenesis is limited and controversial. It is not fully understood if there exists a common factor between all EAEC that contributes to its shared adherence phenotype. EAEC adhere to HEp-2 cells and to each other in a "stacked-brick" configuration, thereby creating a thick biofilm of bacteria that adhere

loosely to the mucosal surface. EAEC secretes enterotoxins and cytotoxins and causes mild but severe mucosal damage and watery diarrhea. No single virulence factor has conclusively been associated with EAEC virulence.

EIEC/*Shigella* (will just be referred to as EIEC) is the only pathovar that are truly invasive and penetrates the epithelial cells. EIEC is also distinguishable from other pathovars due to its lack of flagella or adherence factors. EIEC carries a plasmid that encodes a T3SS that is used to secrete a number of proteins involved in invasion of host cells, including cell uptake, lysis of the endocytic vacuole and apoptosis of macrophages. Infection commences in the colon where the bacteria are transported to the submucosa layer, through microfold cells (M cells). Bacterial cells then go through macrophage uptake that initializes cell death in the macrophage that ultimately leads to release of the bacteria. The EIEC in the submucosa layer invades colonocytes through the basolateral side. Following invasion of colonocytes, the bacteria hijack the host machinery to prevent detection from the immune system and spread to neighboring colonocytes.

DAEC creates a diffuse adherence pattern on HeLa and HEp-2 cells induced by a group of adhesins collectively known as Afa-Dr adhesins. All DAEC establishes attachment to epithelial cells through binding of the receptor decay-accelerating factor (DAF). This binding leads to up regulation of DAF receptors on the apical side of the epithelial cells and provides tighter attachment to the bacteria. The interaction between Afa-Dr adhesins and the host cell leads to increased levels of Ca^{2+}, which along with other factors is believed to be the cause of diarrhea. Unlike the other pathovars, it is believed that the pathogenesis of DAEC is mainly due to the interactions of the Afa-Dr adhesions and the host cell.

UPEC causes cystitis and acute pyelonephritis in humans. UPEC is not a subgroup of the commensal *E. coli* found in the colon, UPEC isolates contains pathogenicity islands specific to UPEC strains that are not found in fecal *E. coli* strains. A urinary tract infection is likely to start with the colonization of the colon alongside the normal gut flora. The bacteria then ascend the urethra into the bladder. Attachment of the bacteria in the bladder is dependent on an important virulence factor that encodes the fimbrial adhesin FimH. The attachment of the bacteria by FimH triggers an invasion of the host cell. Inside the host cell, the bacteria replicate and form biofilm-like complexes known as intracellular bacterial communities (IBC). Motile bacteria leave the epithelial cell and enter the lumen of the bladder. The attachment and infection of UPEC leads to apoptosis and exfoliation. Some UPEC strains can further ascend to the kidney, this requires that the bacteria "turns off" the fimbriae which then leads to decreased attachment and increased motility due to an increased level of flagellated bacteria. Upon reaching the kidney the attachment to renal cells might be dependent on the expression of P fimbria, although this correlation is still inconclusive.

MNEC is an interesting pathovar because its invasion of the central nervous system offers no apparent advantages to the bacteria. It is very likely that the virulence factors causing disease in humans have in fact been adapted for another purpose [3]. After the initial colonization of MNEC the bacteria is transported by transcytosis through enterocytes into the bloodstream. In the bloodstream the bacteria needs to immediately protect itself against the innate immune system. An antiphagocytic capsule provides protection from the host immune response. MNEC has also been shown to invade macrophages and monocytes. MNEC is transported through the bloodstream to the brain microvascular endothelial cells, where FimH and OmpA

mediate attachment. Ultimately this attachment leads to MNEC crossing the blood-brain barrier and causing edema, inflammation and neural damage.

Typing of Escherichia coli

To type a bacteria is to associate the bacteria with a unique label that can genotypically and/or phenotypically distinguish the specific type of bacteria from other bacteria. Typing bacteria enables the description of bacteria transmission routes and dissemination. Typing is therefore a very important tool in epidemiology and is among others used in outbreak investigations, disease surveillance, and bacterial population studies. Typing is done on the subspecies level, which is why typing is sometimes referred to as sub-typing. Several different strains can belong to the same type but it can also happen that one strain can consist of several types. The latter will mostly be the case in pandemics or long-term evolution studies.

E. coli is especially causing outbreaks related to contaminated food and water and in these situations it is crucial to track down the source of the outbreak as fast as possible. First, the ongoing outbreak needs to be detected and then the source needs to be located. In both these steps, typing is needed and the speed and success of the outbreak investigation is highly dependent on the typing methods employed. The usefulness of a typing method can generally be put in to two categories: performance and convenience [25]. Performance covers typeability, stability, discriminatory power, epidemiological concordance, and reproducibility. Convenience covers: rapidity, flexibility, accessibility, ease of use, and cost. The number of different strains that can be typed with a method defines typeability of the method. Stability covers the stability of the markers, which is employed by a typing method. Discriminatory power is a methods ability to differentiate different unrelated strains randomly sampled from a population. The results of a method should reflect the epidemiological data and it is defined as the epidemiologic concordance of the method. The last of the performance measures is reproducibility. The typing results

should be identical, independently of the time and place where they are obtained. A typing method that performs well must also be convenient in order to make it practically useful in outbreak investigations and in the clinics and hospitals. As already mentioned the method needs to be rapid. Furthermore it needs to be flexible with regard to the number of species that it is able to type. The accessibility of reagents and people with the proper skills is important. From this follows also "ease of use". Does the method require a lot of labor and interpretation of results? Finally, one of the most important aspects for many institutions: cost. With a significant throughput of bacteria that needs typing at reference laboratories and hospitals, the cost must be an important consideration. The most widely used typing methods for *E. coli* will be discussed briefly below.

Serotyping

Probably the most important typing method for *E. coli* is, and has been for quite some time, serotyping [25]. O typing and H typing defines an *E. coli* serotype. K typing is sometimes also employed. The general idea is that specific sera react to a specific antigen and it is this reaction that is observed in the laboratory. O typing refers to cell wall antigens, H typing refers to flagella antigens, and K typing refers to capsule antigens [26,27]. Not all *E. coli* are motile and are therefore not H typed. Some *E. coli* cannot be O typed, although in general these are a minority, these are normally referred to as "O rough". Interestingly, most non-motile *E. coli* can be H typed by whole genome sequencing (WGS) and all *E. coli* seems to be O typeable by WGS (data not shown). But even conventional phenotypic

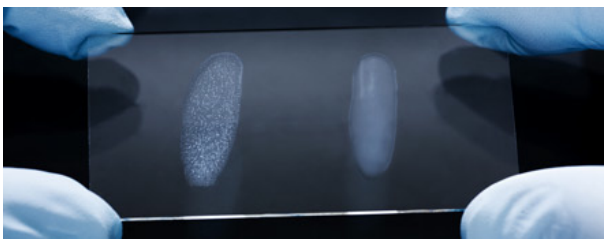

Figure 2. Serotyping. Positive reaction to sera (left) and negative reaction (right). *Figure is borrowed from www.ssi.dk*.

serotyping has high typeability. The genes responsible for serotype are involved in virulence and are therefore under selective pressure from the host immune system. From years of experience, serotyping has proven to be quite stable, although variable enough to provide significant discriminatory power and epidemiological concordance. If serotyping is done very carefully and under standardized conditions it is also reproducible. In general the performance of serotyping is quite good. However, the level of convenience is not as high. Serotyping is not especially fast, it takes at least 3 days for the O and H typing [28]. Serotyping is species specific but the principle can be transferred to other species and has so with for example *Salmonella*. As mentioned it is important that serotyping is done in a systematic and standardized way in order to create confident results but even then, the results are not clear and it requires experience and skill to tell positive reactions from negative ones. Furthermore the method requires sera that are not cheap due to the rather laborious production process, involving live animals. The cost of consumables, manpower and time makes serotyping expensive [18]. It should also be kept in mind that although serotyping provides good discriminatory power, the typing does not provide any information on the relationships between the different serotypes.

Pulse Field Gel Electrophorese (PFGE)

PFGE is a molecular typing method where the genome of the bacteria is fragmented using restriction enzymes. The restriction enzymes utilized are so-called “rare cutters” meaning they usually yield fewer than 30 fragments. The fragments are then run on a gel with an electric field which angle is changed periodically, in order to allow the large fragments to move through the gel. The fragments will move through the gel according to size and create a collection of bands that defines the type, almost like a barcode. The type of the isolate is then defined based on a specific set of criteria [29].

PFGE has high typeability since all strains can be typed. The stability is difficult to evaluate since the *E. coli* genome is known to be quite dynamic, however far from all genomic changes will result in a different band pattern. The discriminatory power of PFGE is very high and apart from WGS probably the highest of all available typing methods for *E. coli*. PFGE shows epidemiological concordance although band patterns for a specific strain can change during an outbreak. Efforts have been done to describe relationships between different band patterns, but such evaluations should be done with much care, since almost identical patterns can be obtained from two unrelated strains and very different patterns can be obtained from related strains. Through careful standardization PFGE is reproducible. As with serotyping the performance of PFGE is quite good but the level of convenience might be as low if not lower than serotyping. PFGE takes at least 2-4 days [25,28]. PFGE can easily be applied to other species, although the restriction enzymes employed might differ. The PFGE method requires relatively expensive equipment and skilled personal. Even though computational software exists to analyze and process the gel images, it still requires experienced personal to carefully evaluate the gels. The similarity of PFGE band patterns should not be considered a measure for genetic distance [25] and band pattern similarity can not be guaranteed to tell anything about strain similarity. However, in practice band pattern similarity are used to infer relationship between strains [25].

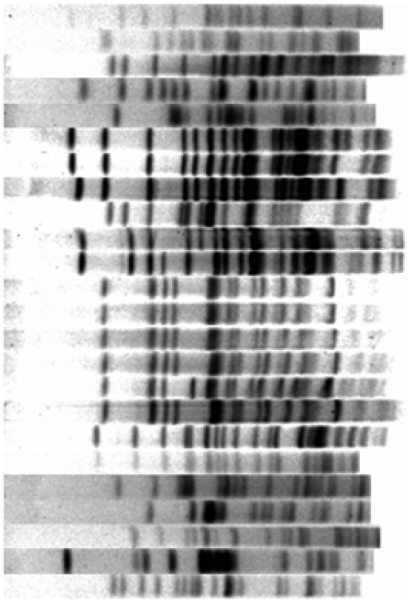

Figure 3. PFGE of 24 *E. coli* strains isolated from pigs. *Figure is a cropped version from Blanco et al.* [80].

Multi Locus Sequence Typing (MLST)

MLST is based on the sequencing of selected conserved housekeeping genes. Three MLST schemes currently exist for *E. coli*, each with three different sets of genes. Mark Achtman's scheme was the first to be published and contains 7 genes. The Pasteur institute published a second scheme consisting of 8 genes. The third scheme was developed specifically for shiga-toxin producing *E. coli* (STEC/VTEC) and contains 15 genes, although a type can consist of 2, 7, and 15 alleles. Regardless of scheme, the approach is similar. The genes specified by the scheme are sequenced for the isolate that needs to be typed and the allele is compared to a database of known alleles. The specific combination of alleles then constitutes the type that is indicated by a single number. For example most O157:H7 isolates are sequence type ST11. The typeability is in principle high, since the housekeeping genes can be found in all *E. coli* isolates. However, alleles that are not already known or combination of alleles that hasn't been seen before cannot be assigned a specific type. It obviously provides validity to the MLST database that all types are manually curated, but it is also a considerable drawback. The stability of the MLST genes is quite high, since these genes are essential for bacterial survival. The stability is in fact so high that it negatively influences the discriminatory power of this typing method, which is significantly lower than both serotyping and PFGE. An estimation of the variation in *E. coli* genes also suggests that the genes used for MLST typing contains less variation than the average conserved gene found in *E. coli*, with the exception of the Pasteur scheme that actually showed to contain genes more variable than the average conserved gene [**Kaas I**]. MLST typing exhibit limited epidemiological concordance due to the low discriminatory power, non-outbreak isolates will have a significant chance of being clustered with outbreak strains due to identical sequence type.

Inferring phylogeny based solely on MLST sequences infers incorrect relationships between strains [8]. Inferring a phylogeny based on the conserved genes of 186 *E. coli* isolates does suggest that the actual assignment of sequence types does infer a clonal relationship between the assigned strains [**Kaas I**]. Reproducibility is high and the very simple assignment and globally shared databases makes comparison of different types very simple. MLST typing is like the previously mentioned methods not very convenient. The typing method is slow, due to the individual sequencing of at least 7 different genes, although automated systems do exist that can do MLST typing in a single day [30], most laboratories don't have this option. Each species needs its own MLST scheme, but the principles behind MLST typing apply to all species. As mentioned it will be quite laborious for most laboratories to do sequencing of 7 specific genes and the overall cost is substantial. Traditional MLST typing is outdated due to the vast improvements to WGS. It has become cheaper to sequence the entire genome than just 7 specific genes [31]. At the time of writing new sequence types are only accepted through manual curation of Sanger sequencing. With fewer and fewer people doing Sanger sequencing new curation methods that can handle WGS data needs to be implemented if new types are to be added.

Next generation sequencing (NGS) in epidemiology

Next generation sequencing has made WGS widely available. WGS is currently cheaper than ever, and with the release of several benchtop sequencers, even smaller laboratories are able to do sequencing locally. The vast increase in WGS has also meant a huge increase in the demand for bioinformatics, a relatively new field that with NGS has gotten a lot of attention. The use of Single Nucleotide Polymorphisms (SNP) to infer phylogenies and thereby predict transmission routes in outbreaks have received a lot of attention due to impressive results in several studies, for example a

study on the spread of methicillin resistant *Staphylococcus aureus* (MRSA)[32]. Inferring phylogenies by SNPs has almost become the golden standard for describing outbreaks and has proven its worth in several additional species like *Vibrio cholera* [33], *Mycobacterium tuberculosis* [34], *Salmonella* [35], and also several examples of *E. coli* [36,37] [**Kaas IV**] (SNP typing will be discussed in more detail later). However, all these studies have been done retrospectively and while they have helped shed light on how the outbreaks spread, none of them have actually helped to decrease the number of infections. WGS needs to be employed in the front line offices of doctors and epidemiologists and applied in real-time. A pilot study done by Joensen and colleagues has suggested that it is possible for WGS to compete in both cost and speed with the conventional typing methods in Denmark [28]. Bioinformatics has been and continues to be a bottleneck in many projects, but the publication of several freely available bioinformatic tools will hopefully help to make the bottleneck wider. It is now possible to do species identification [38], MLST [39], find resistance genes [31] plus *E. coli* virulence genes [28], SNP calling plus phylogeny [**Kaas III+IV**], and very soon serotyping of *E. coli* (tool is validated and working, manuscript is being composed by Joensen et. al) from raw sequencing data using freely available web-tools.

The ambition with these new possibilities should not be to just replace conventional typing methods. The development of a whole genome typing scheme that could easily be applied, compared and stored in an international database would facilitate the possibility for real-time surveillance and detection of pathogens, virulence genes/plasmids, and resistance on a global scale. Such surveillance and detection is increasingly valuable due to the globalization. Contaminated foods can easily travel all over the world, as can asymptomatic carriers of pathogens. It has always been an

issue in epidemiology that the registered infections only make up a small part of the actual spread of a pathogen. A lot of infections are handled without medical assistance and a lot of infections are treated by broad antibiotics and therefore never actually identified (typed). Furthermore, asymptomatic carriers are seldom found unless they are suspected sources of an outbreak. It has therefore been hypothesized that metagenomic sequencing of certain "hotspots" like sewers or wastewater plants might improve surveillance, because in these samples also the organisms shed in the feces of healthy people are picked up. Apart from the scientific and technical challenges posed by such ambitions, also political and ethical issues need to be dealt with. Fortunately international initiatives like the Global Microbial Identifier (GMI) has been started and hopefully will continue to find support and resources because these initiatives has the potential to have a massive impact on global public health.

Defining a gene

In order to use WGS to its full potential, an important issue is the ability to define genes. The main reason for defining genes are to predict gene function and from that infer theories about the organism in which they reside. Another reason for defining genes is to track the specific evolution of specific genes or collections of genes. For the latter it is critical that sequences defined, as one gene is not mixed with sequences from another gene since that would mix different phylogenetic signals. This can be particularly challenging if paralogues or co-evolution is involved.

Gregor Mendel was the father of modern genetics, although he actually never used the term "gene". He did nonetheless describe discrete recessive and dominant traits/characters (German: Merkmal) that got transferred from parent to offspring. Mendel described how genotypes affected the phenotypes of pea plants in his experiments, without actually knowing about genotypes and phenotypes [40]. It was the Danish botanist Wilhelm Johannsen who coined the word "gene" in 1903 [41] and later "phenotype" and "genotype" in 1909 [42]. Today we define genes as stretches of DNA or RNA that encodes polypeptides or RNA chains that in turn provides a function in the organism. The same gene in different organisms, if expressed, will provide the same function. However, a specific gene can due to mutation or recombination be found in different variants, named alleles. This means that different variants of the same gene have different DNA sequences. It is this theory that scientists rely on when they annotate function to a gene based on its sequence similarity to another annotated gene. Homolog sequences are expected to have identical functions.

One of the most obvious questions in defining genes is probably: How variable can sequences be and still retain the same function? There is probably not a single answer

to this question, because it could vary between species and it certainly varies between conserved genes and accessory genes. Here we define accessory genes, as genes not found in all strains of the same species. All genes found in a specific species make up the species pangenome. The coregenome is considered to be all the genes conserved in strains of a specific species. It is expected that virulence genes like adhesins could be very different and still retain identical function while ribosomal genes can't be too variable. Different attempts have been made to cluster gene sequences. Tettelin *et al.* aligned all gene sequences of a specific isolate to all the gene sequences of another isolate, using three different alignment methods, if any of the three methods found a match with more than 50% nucleotide/amino acid identity and covering at least 50% of the gene/protein length, then the genes matching would be considered to have similar function [43]. Rasko et al. employed a BLAST score ratio (BSR). The BSR is a normalized BLAST value and was employed alongside a more stringent threshold of ~80% identity over the length of the protein. The exact method cannot be deducted from their publication [44]. Reciprocal BLAST is a method where a hit is only considered if the top hit from BLASTing a gene from genome A against genome B is identical to the top hit obtained from BLASTing that gene from genome B against genome A. Touchon et al. applied reciprocal BLAST in order to define orthologous gene clusters in *E. coli*. Genes were considered orthologous if the reciprocal best hit was unique and the amino acid identity was at least ~85% and less than 20% percent different in protein length [45]. Snipen and Ussery suggested a method similar to Touchon's for creating pangenome trees, although they employed the 50% identity and 50% length thresholds as suggested by Tettelin et al. [46]. Lukjancenko et al. employed the Snipen and Ussery method on 61 *E. coli* genomes [8]. The latter method was applied to the 186 genomes used in [**Kaas I**]. Some of the gene clusters

obtained was relatively large and on closer inspection it was clear that a lot of unrelated gene sequences had been clustered together (data not shown). The issue encountered with this method is likely to cause an issue in every one of the previously mentioned gene clustering methods. It was found that large gene clusters would start to act as “black holes” and eventually all genes would be part of the same gene cluster. The problem is caused by the increasing variety allowed between the genes in the cluster and the newly ones that gets added. As more genes are added to the cluster the greater the distance between the initial genes added can be to the newly added genes. The gene clusters has most likely been too small to act as “black holes” in the previous studies where a much smaller number of genomes were analyzed. More complex clustering was needed for the [**Kaas I**] study. In 2003 before any of the previous mentioned methods was published, Li et al. actually developed a method to cluster orthologous genes named orthoMCL [47]. As the name suggests this method uses the Markov Chain Clustering (MCL) algorithm developed by Van Dongen [48]. In [**Kaas I**] BLAT [49] was applied to create the all-against-all protein alignments and from these a graph was created and used as input to the MCL algorithm. This method is very similar to the orthoMCL method. The clusters created with MCL did not seem to create “black holes” and the bioinformatic tests applied to the clusters suggested that the clusters was well defined [**Kaas I**].

The *E. coli* genome

It is widely accepted that the *E. coli* genome in general is diverse, dynamic and exhibits chromosomal plasticity. The rate of lateral gene transfer is high and the gene residency is relatively short, resulting in a high gene flux in *E. coli* genomes [45]. The high gene flux is also reflected in phylogenies expressed by pangenome trees created from *E. coli* proteomes. Only the phylogeny of highly related genomes can be inferred reliably with this method because only the most recent gene transfers will reflect the true phylogeny [45][**Kaas I**]. Touchon et al. showed that recombination events contribute more to the diversity in the *E. coli* genome than does mutation events. It was further shown that the high gene flux is compatible with the relative clonal nature of *E. coli* due to gene acquisitions and losses primarily happening at certain "hotspots" on the *E. coli* genome, thereby retaining its core genes and also the organization of these [45]. The strong phylogenetic signal among the core genes was also found within a substantial larger amount of genomes (186 genomes, Figure 4) [**Kaas I**]. The clonal nature of *E. coli* was observed using Multi Locus Enzyme Electrophorese (MLEE) that defined four main groups (A, B1, B2, and D) into which *E. coli* can be divided [50,51]. Two accessory groups were further defined (C and E) [5,52], and later a sub-group "F" of D was defined [53]. The overall phylogenies created for *E. coli* generally agree that the first split in *E. coli* history lead to the emergence of B2. This is further supported by the increased diversity in this group, which gives it sub-species characteristics. Group D was then the next to emerge, followed by E and lastly A and B1 that are classified as sister groups [5,45] [**Kaas I**].

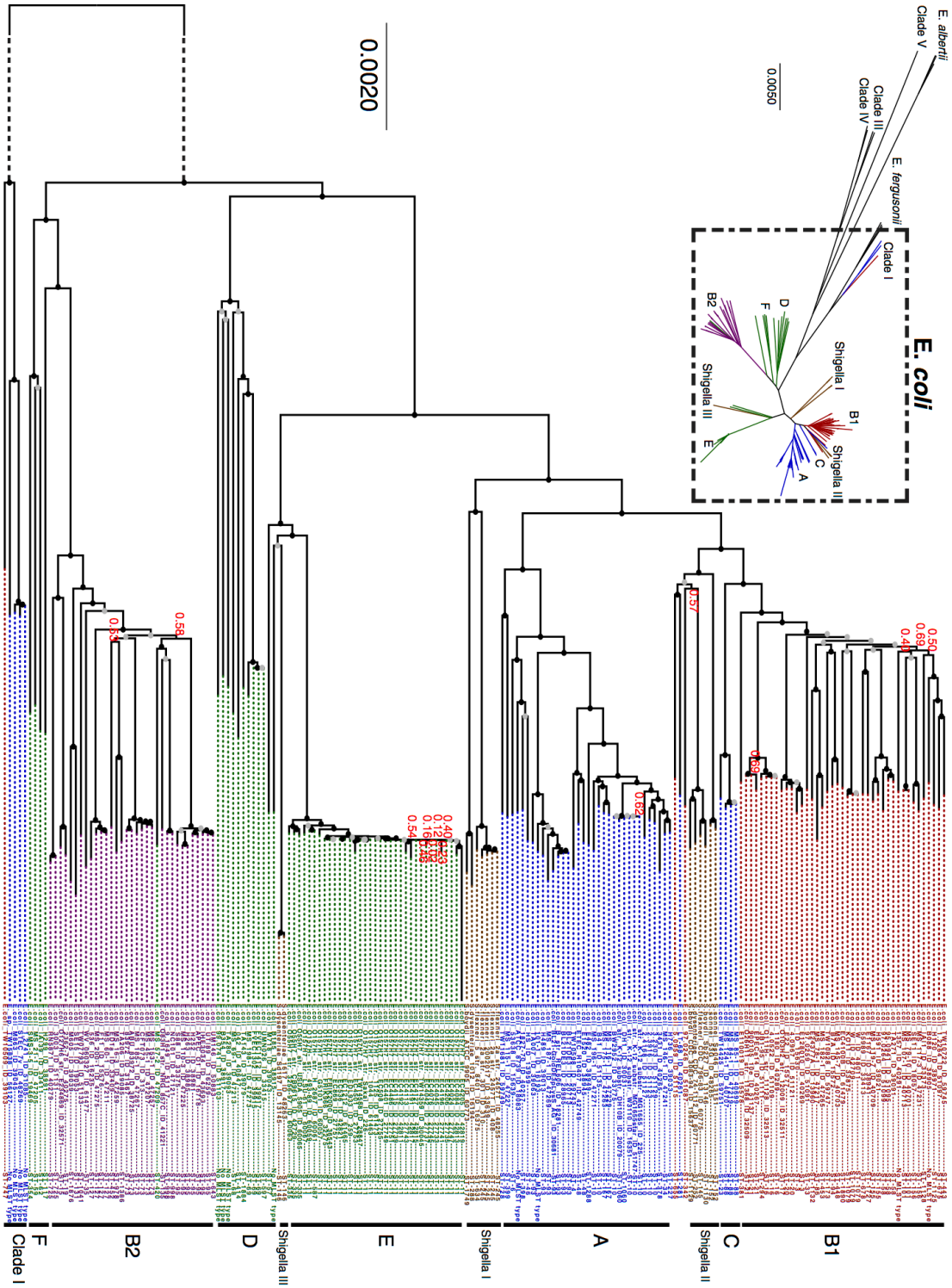

Figure 4. *E. coli* core gene tree. The E. coli tree was created from the alignment core-genes from the 186 E. coli genomes. MLST types are annotated to the far right of each genome name. The Escherichia genus tree was created from 297 core-genes. The phylotypes are marked with the colors blue (A), red (B1), purple (B2), green (D), and the Shigella genomes are marked with the color brown. At each node a black circle indicates a bootstrap value of 1, a grey circle a bootstrap value between 1 and 0.7 and a red number indicate an actual bootstrap value below 0.7. The dashed line in the figure represents a branch, which has been manually shortened by the authors to fit the figure on a printed page.

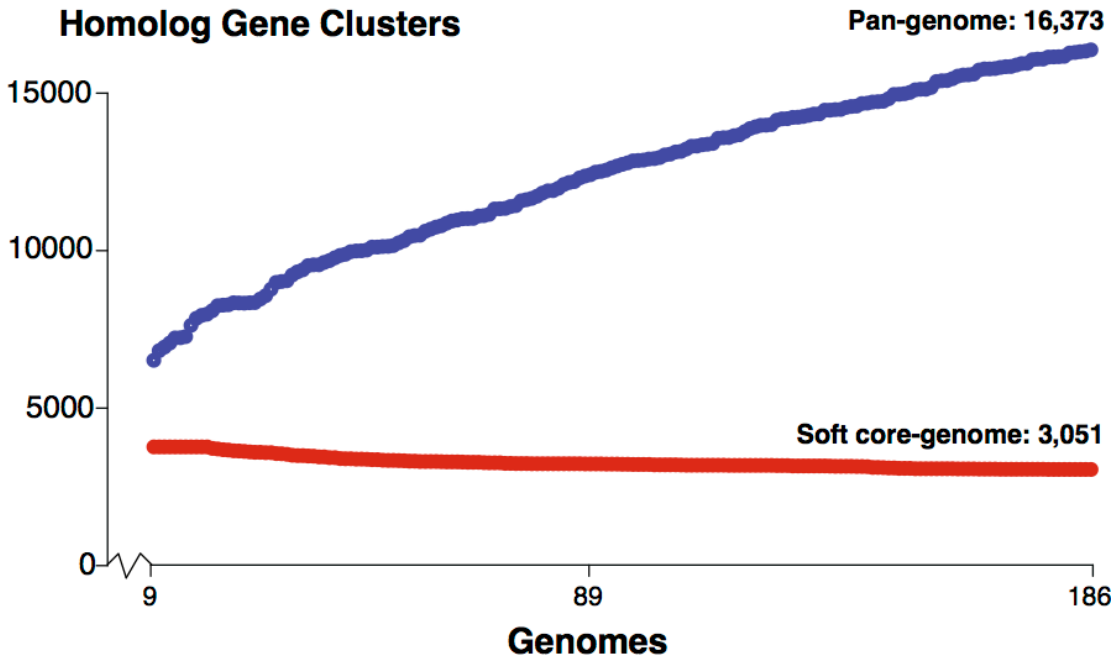

Figure 5. Progress of Homolog Gene Cluster calculation in Kaas I as each genome is added. Two circles exist (red & blue) for each genome added from genome no. 9 up to and including genome no. 186. Red represents the number of core genes after the addition of a genome and blue represents the number of pan-genome genes after the addition of a genome.

Several studies have estimated the core genome of *E. coli* to be around 1,500-2,000 genes [8,45,54–56][**Kaas I**,]. Touchon et al. hypothesized that the genome of the most common ancestor of *E. coli* predicted in their study containing 4,043 genes might better represent the true essential genes of natural living *E. coli* than the core genome because the core genome found in their study lacked 23 high-confidence essential genes [45]. It was these results that inspired a "soft core" genome in which core genes had to be found in 95% of all the analyzed genomes in contrast to the usual 100%, this leads to a soft core genome of 3051 genes [**Kaas I**]. The pangenome has also been estimated in several studies and generally follows the rule that more genomes equal a larger pangenome. Touchon *et al.* found a pan-genome of 11432 genes among 20 genomes [45] and a pangenome of 16676 genes was found among 186 genomes in [**Kaas I**]. A study by Snipen et al. estimated the pangenome of *E. coli* to be around 45000 genes [57]. Lapierre *et al*. speculate that bacterial pangenomes are "open" and therefore will keep increasing as bacteria evolve [58].

Whole genome typing

The entire DNA sequence of an organism can now be made available and thus provide researchers, epidemiologists, and doctors the ultimate material for typing. Indeed, every single DNA sequence is in principle a type of its own, although somewhat impractical. Typing based on WGS involves some level of clustering of the DNA sequences. As mentioned previously, probably the most popular is based on Single Nucleotide Polymorphism (SNP), alternatives will also be discussed.

Single Nucleotide Polymorphism (SNP) analysis

The SNP analysis consists of finding/calling SNPs in the organisms in question and based on these infer a probable phylogeny. The method is based on the theory that random single mutations will happen independently over time throughout the genome of an organism. The amount of SNP differences between two organisms will define the genomic distance between them and in turn define their relationship.

In order to define differences, one needs a common reference to which these differences can be defined. This reference is often the genome of a specific isolate. The DNA sequence of the isolates in question is mapped to the reference genome and the differences are found. This is referred to as “SNP calling”. A SNP is defined by its position in the reference genome and the nucleotide shift. SNP calls are usually stored in a Variant Call Format (VCF). Several tools exists for mapping raw sequence data to a reference sequence [59,60] or mapping assembled genomes to a reference [61]. Likewise several tools also exists for SNP calling [62–64]. However, there seems to be a gap in the availability of bioinformatic tools that will progress the analysis from SNP calls to an inferred phylogeny. A free web-tool “snpTree” was developed and published to fill this gap [**Kaas II**]. The reason for this gap might have something to do with some of the unpleasant assumptions one needs to make at this step, which

might also explain why this step is often not described even high impact publications [32]. In [**Kaas II**] it is assumed that all mutations are independent. Therefore all SNPs can be concatenated in to one single alignment. A difficult issue is how to deal with "missing" SNPs. In [**Kaas II**] the rather large assumption is made that if a SNP is found in one isolate and not in another it is because the other isolates are identical to the reference at the position in question. The assumption of independent mutations is made in all SNP analysis, most analysis also creates a SNP alignment, but a distance matrix could be created instead of an alignment. From the alignment/matrix a phylogeny can then be inferred, snpTree uses Maximum likelihood but many methods exist, including Maximum parsimony and UPGMA.

The assumption with regard to "missing" SNPs made by snpTree is fair as long as the genomes compared (including the reference genome) are closely related and horizontal gene transfer is at a minimum [**Kaas II**]. However, the assumption becomes problematic if the reference isn't closely related or if the sequence data has been obtained by different sequencing methods [**Kaas III**]. Issues with the relatedness of the reference come from the obvious fact that not all "missing" SNPs are caused by identical sequence. A SNP can be "missing" because the actual sequence in the reference does not exist in the genome in question. A SNP can also be missing due to the lack of confident base calls in the specific position, so there might actually be a SNP but it just can't be called with confidence. The assumption made by snpTree will make genomes seem more related to the reference sequence than they actually are. Furthermore, even if sequences are closely related, snpTree will fail if the quality of the sequence data isn't high enough [28].

Different sequencing methods have been shown to contain systematic biases [[65–70]]. Traditionally, studies have simply coped with this by using just a single

sequencing method, although sometimes including assembled sequences. Adding validation of all positions included in a SNP analysis and thereby not making the "missing SNP" assumption, resulted in a much more robust method that was also capable of handling raw sequence data obtained by several different methods [**Kaas III**] (Figure 6). The new method has also been published as a freely available web tool: "Call SNPs and Infer Phylogeny (CSI Phylogeny)" (http://cge.cbs.dtu.dk/services/CSIPhylogeny/).

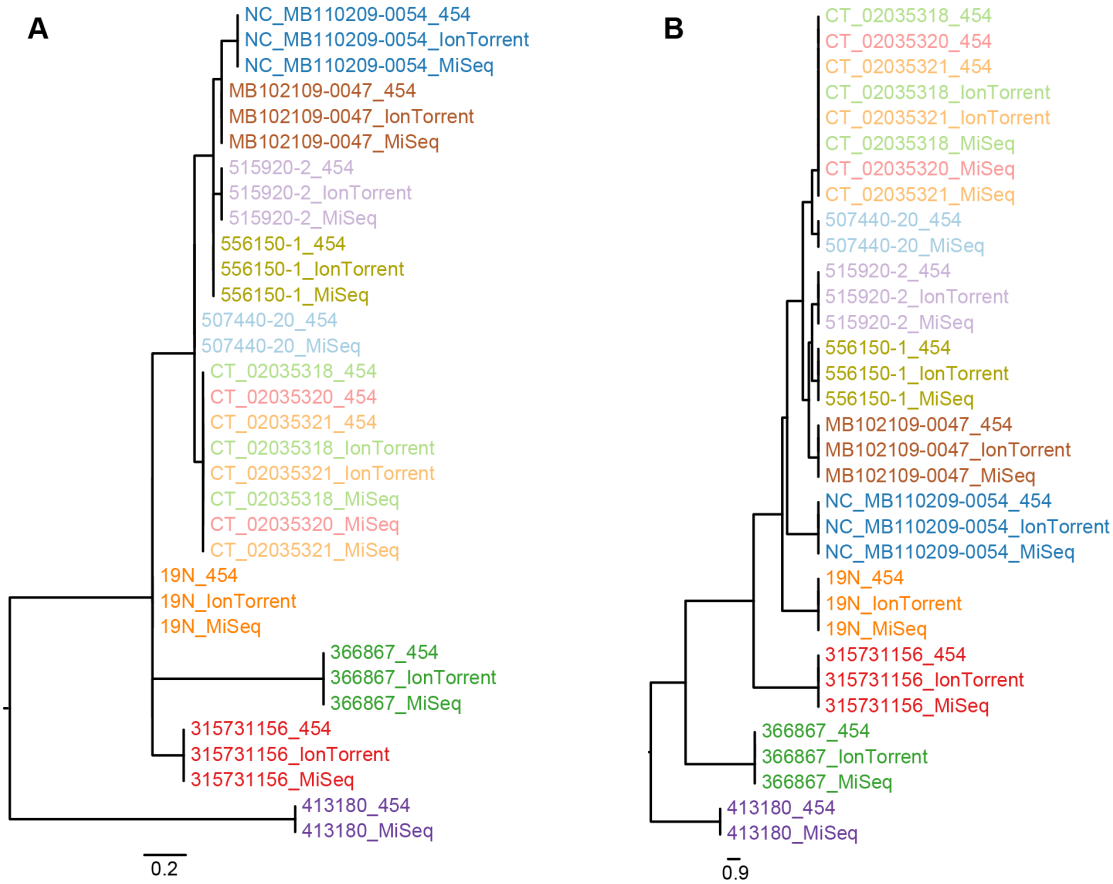

Figure 6. *Salmonella* Montevideo phylogeny based on sequence data obtained from different technologies. Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. (A) Phylogeny inferred with novel SNP procedure; (B) Phylogeny inferred with the Nucleotide Difference (ND) method.

An issue that was also brushed upon earlier is that of horizontal gene transfer, phages, and selfish DNA. Here we will refer to all of these under one common term: "mobile elements". The mobile elements of bacterial genomes challenge the assumption of single independent random mutations. The mutations located on mobile elements are

not independent as they “travel” with the mobile element. The insertion or deletion of an element will also add or delete several mutations, respectively. One mutation is no longer equal to one evolutionary event and this can disturb the true phylogenetic signal. The simplest way of dealing with mobile elements in SNP analysis is to “prune”. Simply ignore all SNPs that are found in close vicinity of each other. This method relies on the fact that mobile elements are often non-essential and are therefore more likely to fix mutations, and if a mobile element maps to another mobile element this will cause an unusual increase in SNP calls in that area [**Kaas II, III**]. Another way of attempting to handle mobile elements is to only call SNPs in the core genome [71] or only call SNPs at specific “trusted” positions. By using known mobile elements, one can also try and locate these by alignment (ex. BLAST) and exclude SNPs found in these regions in the reference [45]. Finally, by annotating the reference genome and “manually” excluding the parts that is believed to disturb the true phylogenetic signal [72]. Using a core genome or pruning is probably the most objective methods, but might also be too simplistic. Finding known elements obviously requires a database of known elements and doesn’t exclude new/unknown mobile elements. Manual curation has the benefit of complete control but with that comes also the risk of the biggest bias. In a study Price and colleagues removed just a single mobile element and documented the exact location, making reproduction of their results easy [72]. However another study published by Gardy and colleagues provides no description on what has been removed or at which locations, making their results impossible to reproduce and difficult to asses the validity of the exclusions [73]. Apart from removal of mobile elements, SNP analyses also contain a wealth of parameters that can be adjusted, mostly with regard to filtering of SNPs. All these issues makes standardization of SNP analyses difficult, but the main obstacle towards

standardizing is the complete dependence on a specific reference. SNP calls made in two different isolates cannot be compared unless the SNPs were called using the exact same genomic reference sequence. Although a method has been published that does not rely on a traditional reference [74], it still relies on a computation that cannot be compared across studies. In its essence, the method takes raw sequence data from isolates and finds the SNPs located in the core genome of these isolates. It remains to be seen if this method is an improvement to regular de novo assembly followed by SNP calling by MUMMER for example. However, the greatest issue with the method is that it only handles data obtained by one sequencing method at a time and including new sequences requires a complete recalculation. It should be mentioned that the method was developed with neither phylogeny nor epidemiology in mind but to find biomarkers in endangered animals.

K-mer, nucleotide difference (ND), and gene-by-gene

There are numerous alternatives to SNP analysis. One alternative is "k-mer phylogenies". These are obtained by fragmenting sequence reads or assembled genomes in to fragments of size "k" (k-mers), counting the k-mers, comparing the counts between isolates, and creating distance matrices [**Kaas IV**]. This method can be implemented to be extremely efficient, but at its present state still contains too much noise in order to resolve clonal outbreaks [**Kaas IV**].

A promising method is the Nucleotide Difference (ND) method published in [71]. This method is processing all the positions in all possible genome pairs and counts the differences between them. The method has high resolution and seems at least as robust as the SNP method "CSI Phylogeny" mentioned earlier, as it also handles sequence data obtained by different sequencing methods [**Kaas III**].

Gene-by-gene is a method that is based on the same principles as MLST [75,76]. The method is (depending on the genes employed) referred to as “Whole Genome MLST (wgMLST)” or “Ribosomal MLST (rMLST)”. The main difference to traditional MLST is that more genes are employed in the analysis. For wgMLST several hundred genes are employed specific to the isolate in question. While rMLST relies on just 53 genes found to be conserved between almost all species [77]. The resolution of rMLST has been shown to be almost comparable to that of SNP analysis [75,77]. The actual computations needed for inferring phylogenies etc. is limited, which is quite positive. It should be noted however, that prior to the application of the gene-by-gene method annotated and assembled sequences are needed. However, the main obstacle for the gene-by-gene approach is without doubt the curation of an allele database. For traditional MLST there exist hundreds of alleles at each locus and several thousand ST types. A gene-by-gene database has been published (PubMLST.org) in an attempt to store alleles for several bacterial species and rMLST. The amount of information in the database at present time is limited and as for traditional MLST typing has an issue with new/unknown alleles. It also seems unfortunate that a database that will rely very much on the participation from the community and its users, although access is open, has restrictions on the use of the data found in the database.

Defining clones

E. coli is responsible for a huge number of sporadic infections but have also caused some of the largest and most severe outbreaks in all parts of the world for example Japan [78], Germany [12] and the United States [37]. These outbreaks could have been limited if it rapidly had been possible to identify the source. It is crucial in the clinical setting and in epidemiology that one is able to distinguish strains and define clones in order to monitor and fight infectious agents. Five methods were evaluated in

[**Kaas IV**]. The five methods were based on SNPs, K-mers, nucleotide differences (ND), core genes, and average nucleotide identity (ANI). The methods were evaluated with regard to resolution and clustering at the outbreak level. It was shown that with just a handful of isolates, all but the ANI method was able to cluster the outbreak isolates into monophyletic clades. It was however only the SNP method that was able to establish a "clone threshold". A threshold of 500 SNPs could distinguish all combinations of a one-to-one comparison between isolates (See Figure 7). It was argued that the limit might be lowered to 200 due to the inclusion of two sets of outbreak strains that would not be considered traditional outbreak strains. The ND method was originally developed for raw data and the results from [**Kaas IV**] suggest that it still needs improvement when dealing with assembled genomes. However, if the assembled genomes are left out of the analysis the ND method obtains comparable results to that of the SNP method – a threshold of about 200 nucleotide differences. Previous *E. coli* studies has recorded up to 74 SNPs within outbreak strains [37]. For *Salmonella*, which is considered a very homogenous species, has been observed within outbreak variation of up to 30 SNPs [71]. An outbreak of *Staphylococcus aureus* in a neonatal unit showed up to 16 SNPs between outbreak strains [79]. These studies confirms the theory that different species evolves at different speeds and therefore clonal definitions will have to be species specific. The study by Kaas *et al.* suggests that *E. coli* isolates having less than 200 SNPs are clones. It seems like a high threshold.

Figure 7. Genetic distance between outbreak and background in 10 different outbreaks. Each outbreak is defined by two bars for each method, a green and a blue one. The name of the outbreak is written below the bottom two bars to which the outbreak belong. Green bars indicate variance found within each outbreak (dark=average, light=max). Blue bars indicate distance to nearest non-outbreak strain. Each blue bar reaching the top expands beyond view. The red bar (horizontal) indicates the clonal threshold for the SNP analysis.

An explanation could be that most of the outbreaks analyzed are food borne outbreaks and therefore are subject to higher variation in contrast to nosocomial outbreaks that might be more homogenous. Another explanation might also be that mobile elements are not completely ignored because the method used for sorting these out was pruning, which might be less efficient than other methods. These results need more data to back up the threshold but the prospects of developing automatic outbreak detection methods are definitely positive.

Future perspectives, challenges & Conclusion

Since the launch of the first next generation sequencer around 2004, sequencing technology has developed at an impressive speed. Cost has been brought down to ensure access to sequence data for even smaller laboratories. The actual machines have recently further undergone adaption to small scale sequencing with the presentation of benchtop sequencers. It is now possible to do sequencing at clinical laboratories and other frontline institutions. Acquiring sequencers, enabling fast local sequencing in the clinical setting and for outbreak investigations should be encouraged as it will most likely free up resources and provide significant health benefits. However, with next generation sequencing also followed a bioinformatic bottleneck. There is a lack of qualified bioinformaticians and a lack of bioinformatic tools available to non-specialists. Researchers has developed a lot of great methods, done fascinating pilot studies, and obtained impressive insights. It does however seem that the focus should be changed from discovery to application. There is a need for studies that evaluates methods with the intend of developing tools and practices that can be applied outside the research community [**Kaas III+IV**][75].

In Denmark the clinical practices send their samples to local microbiological clinical laboratories for analysis. The clinical laboratories then, if the isolate found is among the list of isolates that are under surveillance in Denmark, send the samples to the national reference laboratory ("Statens Serum Institut", SSI). Joensen *et al.* did a study sequencing all VTEC isolates that was received at SSI for several weeks [28]. The pilot study did sequencing along side the classical typing done at SSI and was in a sense "real-time". However, sequencing could be much closer to actual "real-time" - the sampling time, if the sequencing was done at the clinical laboratory. It seems that it would be valuable to do a pilot study where a sequencer is set up at a clinical

laboratory, and do sequencing of everything that would normally be send to SSI. The amount of sequencing could be limited to specific pathogens or geographic location (specific practices) in the beginning. A crucial aspect in such a study would be the direct interaction with the clinical/technical staff. Hopefully the study would elucidate the obstacles in implementing sequencing at clinical laboratories. The study might further enable the creation of tools, databases, and realistic guidelines for future implementations of WGS at other clinical laboratories. If truly successful the results could provide the foundation for a standardized protocol for setting up WGS at a clinical laboratory. Such a study could help close the gap between application and development and help ensure unanimous systems and interfaces.

When clinical laboratories and doctors can use tools on WGS data and see the improved results they will want to participate and in exchange for results provide data that will enable global surveillance of infectious agents including virulence factors and antibiotic resistance. Surveillance on the global scale will provide a vast impact on global health.

Global surveillance will also require handling of huge amounts of data. Even though computer hardware is getting faster, it can't keep up with the amount of resources needed for complex biological analysis or will at least limit it to institutions with significant computational power. The representation of biological sequence in bioinformatics could be improved by implementing libraries that exploits the limited alphabet and create binary code instead of text, as already done to some extend in the programming language Python (https://pythonhosted.org/ngs_plumbing). The binary code requires much less memory and is many times faster to process and compare than text. Binary code coupled with k-mers could potentially create extremely fast phylogenies with a minimum amount of resources.

An area with huge potential is metagenomics, due to the ability to sequence entire communities and the ability to sequence organisms that are not culturable. It is also an area that is full of challenges and its usefulness in practical applications remains to be seen. However, many uses can be imagined. First of all the ability to sequence clinical samples directly will save time and maybe reveal new insights. Sequencing samples from certain "hotspots" such as sewage, in order to do surveillance of pathogens and antibiotic resistance in the community has also been mentioned. There is no doubt that the challenge of small DNA reads from a complex microbiological community mixed together provides significant limitations. New sequencing technologies promises longer reads and in a future perspective, longer reads will significantly increase the usefulness of metagenomics.

Conclusion

The *E. coli* genome is very diverse and has a high gene flux, but retains a strong phylogenetic signal [**Kaas I**] due to the existence of recombination hotspots [45]. In order to make WGS truly useful in outbreak investigations, tools need to be made available to non-bioinformaticians like epidemiologists and clinical laboratories. This includes tools for SNP calling and inferring phylogenies across different sequencing platforms [**Kaas II+III**]. If any bioinformatic method is to be used for whole genome typing of *E. coli*, studies need to be published that challenges the method and that can shed some light on how clones in an outbreak should be distinguished from sporadic cases caused by similar strains [**Kaas IV**].

Whole genome sequence typing of *E. coli* is important, but it will most likely not change anything about the typing nomenclature traditionally used for *E. coli*. Doctors will continue to talk about EHEC, O157:H7, and ST11 types. These are well-

established types, with well-defined properties, and probably don't need to be changed. However, whole genome typing will provide important information on which isolates are clonal and therefore related. WGS will provide information on virulence and resistance, so the name of the organism doesn't really matter, as long as the genome is there. In the end the:

"*The genome is the type.*"

-Ole Lund

References

1. Escherich T (1886) Die darmbakterien des s{ä}uglings und ihre beziehungen zur physiologie der Verdauung. F. Enke.

2. Murray PR, Baron EJ, Pfaller MA, Tenover FC, Yolken RH (1999) Manual of Clinical Microbiology. American Society for Microbiology, Washington DC. EE UU[Links].

3. Hobman JL, Penn CW, Pallen MJ (2007) Laboratory strains of Escherichia coli: model citizens or deceitful delinquents growing old disgracefully? Mol Microbiol 64: 881–885. Available: http://www.ncbi.nlm.nih.gov/pubmed/17501914. Accessed 6 June 2014.

4. Gordon DM, Cowling A (2003) The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. Microbiology 149: 3575–3586. Available: http://mic.sgmjournals.org/cgi/doi/10.1099/mic.0.26486-0. Accessed 10 July 2014.

5. Tenaillon O, Skurnik D, Picard B, Denamur E (2010) The population genetics of commensal Escherichia coli. Nat Rev Microbiol 8: 207–217. Available: http://www.ncbi.nlm.nih.gov/pubmed/20157339. Accessed 19 July 2011.

6. Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos G a, et al. (2009) Cryptic lineages of the genus Escherichia. Appl Environ Microbiol 75: 6534–6544. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2765150&tool=pmcentrez&rendertype=abstract. Accessed 6 March 2012.

7. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, et al. (2011) Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A 108: 7200–7205. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3084108&tool=pmcentrez&rendertype=abstract. Accessed 29 February 2012.

8. Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 Sequenced Escherichia coli Genomes. Microb Ecol 60: 708–720. Available: http://www.ncbi.nlm.nih.gov/pubmed/20623278. Accessed 20 August 2010.

9. Croxen M a, Finlay BB (2010) Molecular mechanisms of Escherichia coli pathogenicity. Nat Rev Microbiol 8: 26–38. Available: http://www.ncbi.nlm.nih.gov/pubmed/19966814. Accessed 28 May 2014.

10. Kaper JB, Nataro JP, Mobley HL (2004) Pathogenic Escherichia coli. Nat Rev Microbiol 2: 123–140. Available: http://www.ncbi.nlm.nih.gov/pubmed/15040260. Accessed 28 May 2014.

11. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, et al. (2009) Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. Proc Natl Acad Sci U S A 106: 17939–17944. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2764950&tool=pmcentrez&rendertype=abstract.

12. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, et al. (2012) Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1121491109. Accessed 7 February 2012.

13. Hebbelstrup Jensen B, Olsen KEP, Struve C, Krogfelt KA, Petersen AM (2014) Epidemiology and Clinical Manifestations of Enteroaggregative Escherichia coli. Clin Microbiol Rev 27: 614–630. Available: http://www.ncbi.nlm.nih.gov/pubmed/24982324. Accessed 21 July 2014.

14. Olesen B, Scheutz F, Andersen RL, Menard M, Boisen N, et al. (2012) Enteroaggregative Escherichia coli O78 : H10 , the Cause of an Outbreak. doi:10.1128/JCM.01909-12.

15. Vieira N, Bates SJ, Solberg OD, Ponce K, Howsmon R, et al. (2007) High prevalence of enteroinvasive Escherichia coli isolated in a remote region of northern coastal Ecuador. Am J Trop Med Hyg 76: 528–533. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2396511&tool=pmcentrez&rendertype=abstract.

16. Kotloff KL, Winickoff JP, Ivanoff B, Clemens JD, Swerdlow DL, et al. (1999) Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. Bull World Health Organ 77: 651–666. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2557719&tool=pmcentrez&rendertype=abstract.

17. Table 3b FoodNet–Incidence of Laboratory–Confirmed Infections by Site 2013 (2014). Foodborne Dis Act Surveill Netw. Available: http://www.cdc.gov/foodnet/data/trends/tables/2013/table3a-b.html#table-3b. Accessed 24 July 2014.

18. Nataro JP, Kaper JB (1998) Diarrheagenic Escherichia coli. Clin Microbiol Rev 11: 142–201. Available: http://cmr.asm.org/content/11/1/142.abstract. Accessed 15 July 2014.

19. Johansen TEB, Cek M, Naber KG, Stratchounski L, Svendsen M V, et al. (2006) Hospital acquired urinary tract infections in urology departments: pathogens, susceptibility and use of antibiotics. Data from the PEP and PEAP-studies. Int J Antimicrob Agents 28 Suppl 1: S91–107. Available: http://www.ncbi.nlm.nih.gov/pubmed/16829052. Accessed 13 September 2014.

20. Jacobsen SM, Stickler DJ, Mobley HLT, Shirtliff ME (2008) Complicated catheter-associated urinary tract infections due to Escherichia coli and Proteus mirabilis. Clin Microbiol Rev 21: 26–59. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2223845&tool=pmcentrez&rendertype=abstract. Accessed 13 July 2014.

21. Bonacorsi S, Bingen E (2005) Molecular epidemiology of Escherichia coli causing neonatal meningitis. Int J Med Microbiol 295: 373–381. Available: http://www.ncbi.nlm.nih.gov/pubmed/16238014. Accessed 24 July 2014.

22. Frankel G, Phillips AD (2008) Attaching effacing Escherichia coli and paradigms of Tir-triggered actin polymerization: getting off the pedestal. Cell Microbiol 10: 549–556. Available: http://www.ncbi.nlm.nih.gov/pubmed/18053003. Accessed 15 July 2014.

23. Tseng M, Fratamico PM, Manning SD, Funk J a (2014) Shiga toxin-producing Escherichia coli in swine: the public health perspective. Anim Health Res Rev 15: 1–13. Available: http://www.ncbi.nlm.nih.gov/pubmed/24397985. Accessed 2 September 2014.

24. Pitari GM, Zingman L V, Hodgson DM, Alekseev a E, Kazerounian S, et al. (2003) Bacterial enterotoxins are associated with resistance to colon cancer. Proc Natl Acad Sci U S A 100: 2695–2699. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=151403&tool=pmcentrez&rendertype=abstract.

25. Van Belkum a, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, et al. (2007) Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect 13 Suppl 3: 1–46. Available: http://www.ncbi.nlm.nih.gov/pubmed/17716294.

26. KAUFFMANN F (1975) CLASSIFICATION OF BACTERIA A REALISTIC SCHEME WITH SPECIAL REFERENCE TO THE CLASSIFICATION OF SALMONELLA-SPP AND ESCHERICHIA-SPP: 169.

27. Orskov F, Orskov I (1984) SEROTYPING OF ESCHERICHIA-COLI. METHODS Microbiol 14: 43–112.

28. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, et al. (2014) Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol 52: 1501–1510. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3993690&tool=pmcentrez&rendertype=abstract. Accessed 27 May 2014.

29. Tenover FC, Arbeit RD, Goering R V, Mickelsen PA, Murray BE, et al. (1995) Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33: 2233–2239. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC228385/. Accessed 7 August 2014.

30. Sullivan CB, Jefferies JMC, Diggle M a, Clarke SC (2006) Automation of MLST using third-generation liquid-handling technology. Mol Biotechnol 32: 219–226. Available: http://www.ncbi.nlm.nih.gov/pubmed/16632888.

31. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, et al. (2013) Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. J Antimicrob Chemother 68: 771–777. Available: http://www.ncbi.nlm.nih.gov/pubmed/23233485. Accessed 25 February 2014.

32. Harris SR, Feil EJ, Holden MTG, Quail M a, Nickerson EK, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327: 469–474. Available: http://www.ncbi.nlm.nih.gov/pubmed/20093474.

33. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, et al. (2011) Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio 2: 1–6. Available: http://mbio.asm.org/content/2/4/e00157-11.short. Accessed 27 February 2014.

34. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, et al. (2013) Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13: 137–146. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3556524&tool=pmcentrez&rendertype=abstract. Accessed 23 February 2014.

35. Allard MW, Luo Y, Strain E, Li C, Keys CE, et al. (2012) High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 13: 32. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3368722&tool=pmcentrez&rendertype=abstract. Accessed 21 February 2014.

36. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, et al. (2013) Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. J Clin Microbiol 51: 232–237. Available:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3536255&tool=pmcentrez&rendertype=abstract. Accessed 17 June 2014.

37. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA (2011) Genomic anatomy of Escherichia coli O157: H7 outbreaks. Proc Natl Acad Sci 108: 20142–20147. Available: http://www.pnas.org/content/108/50/20142.short. Accessed 5 January 2012.

38. Larsen M V, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, et al. (2014) Benchmarking of methods for genomic taxonomy. J Clin Microbiol 52: 1529–1539. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3993634&tool=pmcentrez&rendertype=abstract. Accessed 30 July 2014.

39. Larsen M V, Cosentino S, Rasmussen S, Friis C, Hasman H, et al. (2012) Multilocus Sequence Typing of Total Genome Sequenced Bacteria. J Clin Microbiol 50: 1355–1361. Available: http://www.ncbi.nlm.nih.gov/pubmed/22238442. Accessed 17 March 2012.

40. Mendel G (1866) Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines Brünn 42: 3–47.

41. Johannsen W (1903) Om arvelighed i samfund og i rene linier. Overs over det K Danske Vidensk Selsk Forh 3: 247–270.

42. Johannsen W (1909) Elemente der exakten Erblichkeitslehre.

43. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102: 13950–13955. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1216834&tool=pmcentrez&rendertype=abstract.

44. Rasko D a, Rosovitz MJ, Myers GS a, Mongodin EF, Fricke WF, et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 190: 6881–6893. Available:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2566221&tool=pmcentrez&rendertype=abstract. Accessed 15 July 2014.

45. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5: e1000344. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2617782&tool=pmcentrez&rendertype=abstract. Accessed 21 July 2011.

46. Snipen L, Ussery DW (2010) Standard operating procedure for computing pangenome trees. Stand Genomic Sci 2: 135–141. Available: http://www.standardsingenomics.org/index.php/sigen/article/view/sigs.38923. Accessed 18 September 2010.

47. Li L, Stoeckert JCJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189. Available: http://genome.cshlp.org/content/13/9/2178.short. Accessed 14 May 2012.

48. Dongen S van (2000) Graph Clustering by Flow Simulation University of Utrecht. Available: http://micans.org/mcl/.

49. Kent WJ (2002) BLAT---The BLAST-Like Alignment Tool. Genome Res 12: 656–664. doi:10.1101/gr.229202.

50. Selander RK, Caugant D a, Ochman H, Musser JM, Gilmour MN, et al. (1986) Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol 51: 873–884. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=238981&tool=pmcentrez&rendertype=abstract.

51. Goullet P, Picard B (1989) Comparative electrophoretic polymorphism of esterases and other enzymes in Escherichia coli. J Gen Microbiol 135: 135–143. Available: http://www.ncbi.nlm.nih.gov/pubmed/2674323. Accessed 3 September 2014.

52. Herzer PJ, Inouye S, Inouye M, Whittam TS (1990) Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli. J Bacteriol 172: 6175–6181. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=526797&tool=pmcentrez&rendertype=abstract. Accessed 3 September 2014.

53. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, et al. (2008) Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains. BMC Genomics 9: 560. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2639426&tool=pmcentrez&rendertype=abstract. Accessed 20 July 2011.

54. Vieira G, Sabarly V, Bourguignon P-Y, Durot M, Le Fèvre F, et al. (2011) Core and panmetabolism in Escherichia coli. J Bacteriol 193: 1461–1472. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3067614&tool=pmcentrez&rendertype=abstract. Accessed 18 June 2011.

55. Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW (2007) Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray. Genome Biol 8: R267. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2246269&tool=pmcentrez&rendertype=abstract. Accessed 22 July 2011.

56. Chattopadhyay S, Weissman SJ, Minin VN, Russo T a, Dykhuizen DE, et al. (2009) High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. Proc Natl Acad Sci U S A 106: 12412–12417. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2718352&tool=pmcentrez&rendertype=abstract.

57. Snipen L, Almøy T, Ussery DW (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10: 385. Available:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2907702&tool=pmcentrez&rendertype=abstract. Accessed 11 July 2011.

58. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25: 107–110. Available: http://www.ncbi.nlm.nih.gov/pubmed/19168257.

59. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract. Accessed 7 November 2013.

60. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3322381&tool=pmcentrez&rendertype=abstract. Accessed 10 July 2014.

61. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30: 2478–2483. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=117189&tool=pmcentrez&rendertype=abstract.

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract. Accessed 11 December 2013.

63. DePristo M a, Banks E, Poplin R, Garimella K V, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3083463&tool=pmcentrez&rendertype=abstract. Accessed 9 July 2014.

64. Li R, Li Y, Fang X, Yang H, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19: 1124–1132. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2694485&tool=pmcentrez&rendertype=abstract. Accessed 20 July 2014.

65. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10: R32. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2691003&tool=pmcentrez&rendertype=abstract. Accessed 20 February 2014.

66. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, et al. (2012) Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30: 78–82. Available: http://www.ncbi.nlm.nih.gov/pubmed/22178993. Accessed 27 February 2014.

67. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, et al. (2013) Comparison of sequencing platforms for single nucleotide variant calls in a human sample. PLoS One 8: e55089. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3566181&tool=pmcentrez&rendertype=abstract. Accessed 19 February 2014.

68. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, et al. (2013) Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One 8: e66621. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679043&tool=pmcentrez&rendertype=abstract. Accessed 23 January 2014.

69. Suzuki S, Ono N, Furusawa C, Ying B-W, Yomo T (2011) Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS One 6: e19534. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3096631&tool=pmcentrez&rendertype=abstract. Accessed 27 February 2014.

70. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, et al. (2013) Updating benchtop sequencing performance comparison. Nat Biotechnol 31: 296. Available: http://www.ncbi.nlm.nih.gov/pubmed/23563422.

71. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM (2014) Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica. PLoS One 9: e87991. Available: http://dx.plos.org/10.1371/journal.pone.0087991. Accessed 5 February 2014.

72. Price L, Stegger M, Hasman H, Aziz M, Larsen J (2012) Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock. MBio 3: 1–6. Available: http://mbio.asm.org/content/3/1/e00305-11.short. Accessed 25 May 2012.

73. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, et al. (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364: 730–739. Available: http://www.ncbi.nlm.nih.gov/pubmed/21345102.

74. Ratan A, Zhang Y, Hayes V, Schuster SC, Miller W (2010) Calling SNPs without a reference sequence. BMC Bioinformatics 11: 130. Available: http://www.biomedcentral.com/1471-2105/11/130/. Accessed 9 July 2014.

75. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford S a, et al. (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11: 728–736. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3980634&tool=pmcentrez&rendertype=abstract. Accessed 14 July 2014.

76. Jolley K a, Maiden MCJ (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11: 595. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3004885&tool=pmcentrez&rendertype=abstract. Accessed 18 July 2014.

77. Jolley K a, Bliss CM, Bennett JS, Bratcher HB, Brehony C, et al. (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 158: 1005–1015. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3492749&tool=pmcentrez&rendertype=abstract. Accessed 8 August 2014.

78. Watanabe H, Wada A, Inagaki Y, Itoh K, Tamura K (1996) Outbreaks of enterohaemorrhagic Escherichia coli O157:H7 infection by two different genotype strains in Japan, 1996. Lancet 348: 831–832. Available: http://www.ncbi.nlm.nih.gov/pubmed/8814014. Accessed 9 September 2014.

79. Harris SR, Cartwright EJP, Török ME, Holden MTG, Brown NM, et al. (2013) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis 13: 130–136. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3556525&tool=pmcentrez&rendertype=abstract. Accessed 27 February 2014.

80. Blanco M, Lazo L, Blanco J (2010) Serotypes, virulence genes, and PFGE patterns of enteropathogenic Escherichia coli isolated from Cuban pigs with diarrhea. Int …: 53–60. Available: http://130.206.88.107/index.php/IM/article/view/4c457c86ca5cd.002. Accessed 29 September 2014.

BMC Genomics

RESEARCH ARTICLE Open Access

Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes

Rolf S Kaas[1*], Carsten Friis[1], David W Ussery[2] and Frank M Aarestrup[1]

Abstract

Background: *Escherichia coli* exists in commensal and pathogenic forms. By measuring the variation of individual genes across more than a hundred sequenced genomes, gene variation can be studied in detail, including the number of mutations found for any given gene. This knowledge will be useful for creating better phylogenies, for determination of molecular clocks and for improved typing techniques.

Results: We find 3,051 gene clusters/families present in at least 95% of the genomes and 1,702 gene clusters present in 100% of the genomes. The former 'soft core' of about 3,000 gene families is perhaps more biologically relevant, especially considering that many of these genome sequences are draft quality. The *E. coli* pan-genome for this set of isolates contains 16,373 gene clusters.
A core-gene tree, based on alignment and a pan-genome tree based on gene presence/absence, maps the relatedness of the 186 sequenced *E. coli* genomes. The core-gene tree displays high confidence and divides the *E. coli* strains into the observed MLST type clades and also separates defined phylotypes.

Conclusion: The results of comparing a large and diverse *E. coli* dataset support the theory that reliable and good resolution phylogenies can be inferred from the core-genome. The results further suggest that the resolution at the isolate level may, subsequently be improved by targeting more variable genes. The use of whole genome sequencing will make it possible to eliminate, or at least reduce, the need for several typing steps used in traditional epidemiology.

Keywords: Escherichia coli, Core-genome, Pan-genome, Phylogeny, Whole genome sequencing, Genetic variation, Comparative genomics, MLST typing, Phylotyping

Background

The declining cost of whole genome sequencing (WGS) of bacterial pathogens has now made sequencing an option available for many scientists including those working in routine laboratories. WGS is useful in research and trend studies, but might soon be found in routine applications for diagnostics and surveillance, as well. Depending on the technology, WGS can be done in a few of hours and at low cost. Combined with the right tools, WGS makes real-time surveillance and rapid detection of outbreaks possible [1].

Escherichia coli is a gut commensal bacterium, as well as an important pathogen. As a commensal it acts as a beneficial member of the human microbiome in both digestion and defense against opportunistic pathogens. It is, however, also one of the most important human pathogens as it is responsible for up to 90% of all human urinary tract infections, and a frequent cause of septicemia, gastro-intestinal and other infections. *E. coli* is responsible for a large part of the more than 2 million deaths caused by diarrhea in children under the age of five in developing countries [2]. In developed countries, bacteremia is the 10th most common cause of death and among the Gram-negative bacteria, *E. coli* is responsible for 30% of the cases [3]. Food borne outbreaks are also frequently observed and rapid characterization is important to detect and prevent outbreaks.

* Correspondence: rkmo@food.dtu.dk
[1]DTU Food, The Technical University of Denmark, Kgs Lyngby, Denmark
Full list of author information is available at the end of the article

Pathogenic *E. coli* are traditionally classified on the basis of serotype and/or Multi Locus Sequence Type (MLST). Pulse field gel electrophoresis (PFGE) is also widely used, especially to detect outbreaks, because of its discriminatory power, but both PFGE and serotyping provide little phylogeneticly meaningful information. In contrast, MLST typing often lacks the discriminatory power to describe complex outbreaks [4], but can indicate some phylogenetic relationships, since it is based on the sequencing of genes, although some of these relationships might be questionable [5]. *E. coli* is also classified according to the presence of specific virulence factors in to pathogroups such as VTEC (verocytotoxin producing *Escherichia coli*), ETEC (enterotoxigenic *Escherichia coli*), EIEC (enteroinvasive *Escherichia coli*), EHEC (enterohemorrhagic *Escherichia coli*), EPEC (enteropathogenic *Escherichia coli*) and EAEC (enteroadherent *Escherichia coli*).

Apart from its role in human and animal health and diseases, *E. coli* is also an important and well-characterized model organism, which makes it one of the most sequenced organisms in GenBank, second only to *Staphylococcus aureus* in terms of the number of sequenced genomes available. This makes *E. coli* a good candidate for genome variation studies.

With the application of WGS to epidemiology, the opportunity to create better and more precise typing methods has arisen. To facilitate the future comparison of WGS data and identify clones or related strains, it is important to develop standards for classifying isolates. The genes within a genome are constantly evolving and some genes fix mutations at faster rates than others [6]. This rate is complex because it has several dependencies including gene function, selection pressure and location on the chromosome or plasmid [7].

When choosing appropriate target genes for typing purposes, it is important to know that the targets can be expected to exist in all isolates to be typed. One method for doing this is to choose genes that exist in all members of the species studied – the core-genes.

It is the aim of this study to identify core-genes and to estimate the variation within all the genes of 186 publically available *E. coli* and *Shigella* genomes from GenBank. In addition, different methods for classification of *E. coli* are evaluated. The results form a basis for future implementation of WGS as a standard typing tool for classification of *E. coli* in phylogeny and epidemiology. Standardized classification of bacteria with WGS is crucial if it is to be used in real-time surveillance and quick outbreak detection.

Results

The Prodigal software predicted a total of 945,211 genes across all genomes. This is an average of ~5,082 genes per genome, which could be an overestimation because of the lower quality of some of the draft genome sequences. The average is ~4,837 predicted genes per genome among the complete genomes, which can be compared to the average of ~4,754 genes per genome annotated in the complete genomes in GenBank. The genes were clustered into 16,373 clusters, which represent the *E. coli* "pan-genome". The clusters were determined by MCL clustering, as described in the methods section, and are referred to as Homolog Gene Clusters (HGCs), The "soft core" is defined as all HGCs found in at least 95% of all genomes and the "strict core" is defined as all HGCs found in at least 100% of all genomes. The soft core consists of 3,051 HGCs and the strict core contains 1,702 HGCs.

The progress of the clustering algorithm is plotted in Figure 1. Each point represents the pan- and core-genome results after adding an additional genome. The x-axis starts at genome 9, because each core HGC is allowed to be missing in 9 genomes once each calculation has finished. The size of the core-genome quickly approaches 3,000 HGCs and then stabilizes. The pan-genome continues to rise with the addition of more genomes. The curve seems to become almost linear.

The first 50 added genomes are all complete genomes. There seems to be no unusual drop or rise in the core- or pan-genome, respectively, with the addition of the draft genomes.

Variation within HGCs

The distribution of variation within HGCs is shown in a density plot in Figure 2. The majority of HGCs have less than 0.020 substitutions per site. The 5th and 10th percentiles are also calculated. These show that 95% and 90% of the HGCs have less than 0.242 substitutions per site and 0.179 substitutions per site, respectively.

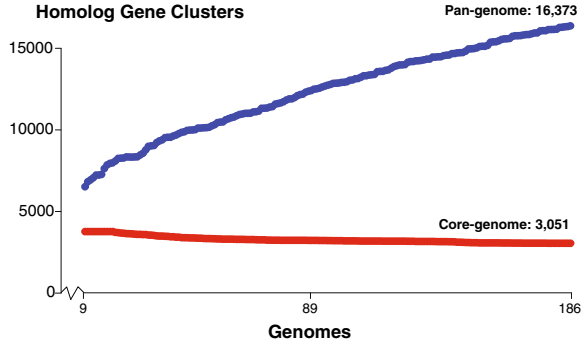

Figure 1 Progress of Homolog Gene Cluster calculation as each genome is added. Two circles exist (red & blue) for each genome added from genome no. 9 up to and including genome no. 186. Red represents the number of core HGCs after the addition of a genome and blue represents the number of pan HGCs after the addition of a genome.

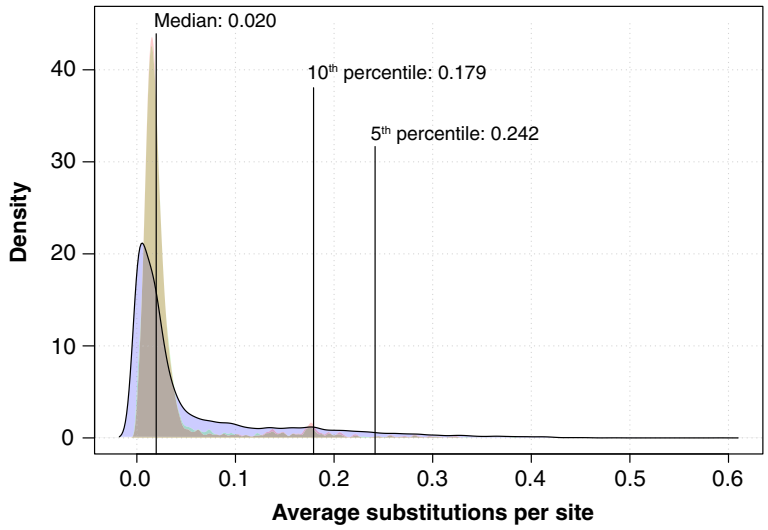

Figure 2 HGC Variation plot. A Density plot was created from the calculation of nucleotide diversity within each HGC. The blue plot was created from all the HGCs. The red plot only includes the strict core HGCs. The green plot includes the soft core (95%) HGCs. Intersection between core plots is yellow.

Nucleotide diversity is calculated as the average number of substitutions per site within an HGC as suggested by Nei & Li [8] (see Materials & Methods for details).

The density plot of the pan-genome (blue) has a single large top, which represents the majority of HGCs. The density plots of the soft core and the strict core are colored green and red, respectively. The intersection of the two cores is colored yellow. It can be observed that the distributions of the two core-genomes are almost identical. The tops of the core distributions are located higher on the x-axis (more diverse), than the top of the pan-genome, but the distributions are narrower, and result in lower medians (~0.018).

1,472 of the HGCs in the pan-genome have zero substitutions per site. This is mostly due to the small sizes of these HGCs; almost half of them contain only two members. One HGC contains 68 members. This HGC represents a small coding sequence of 156 base pairs. It encodes a hypothetical protein named YrhD of unknown function [Swiss-Prot:P58037, EcoGene:EG14370].

The most conserved core HGC was identical for both the soft and the strict cores. It has 188 members (substitutions per site: 0.0000467). Not surprisingly this gene cluster represents a ribosomal gene (S18).

The least conserved soft core HGC has 187 members (substitutions per site: 0.382). It represents a family of conserved genes with unknown function. The least conserved strict core HGC has 1,158 members (substitutions per site: 0.324). It represents a large cluster of ABC transporters. This large family has been reported before, and represents the diverse range of substrate specificities of the different ABC transporters, which is due to substitutions in the periplasmic binding subunit [9].

The least conserved of all the HGCs consists of 28 members (substitutions per site: 0.592). The alignment of this HGC is small and very scattered. It represents a family of transposases. The 28 members only represent 5 different genomes, 3 of which are *Shigella* genomes.

Three distinct MLST schemes exist for *E. coli*, although probably the most widely used is Mark Achtman's set of 7 housekeeping genes (http://mlst.ucc.ie/); the Pasteur institute has created an alternative scheme, which uses 8 genes (http://www.pasteur.fr/recherche/genopole/PF8/mlst/EColi.html), and T. Whittam's scheme uses up to 15 genes (http://www.shigatox.net/) [10-12]. A box plot for the HGCs belonging to each scheme was created and is presented in Figure 3. The genes used in each of the three MLST schemes are presented in Additional file 1. A phylogenetic tree was inferred for a selection of American

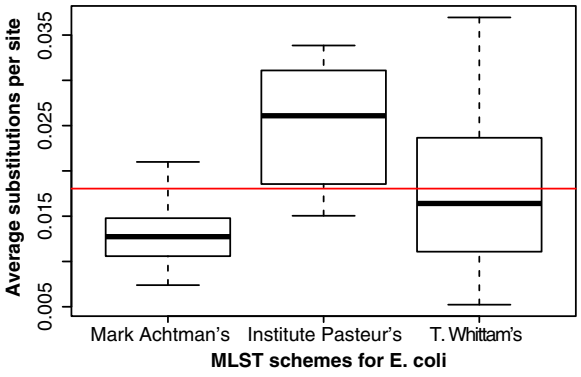

Figure 3 Box plot of MLST gene variation. A box plot presenting the distribution of nucleotide diversity within each of the three MLST schemes. The red line represents the median of percent identity for HGCs in the core (~0.018 substitutions per site).

outbreak isolates with ST type 11 and serotype O157:H7 using the genes from the different MLST schemes. As a proof-of-concept, a phylogenetic tree was also inferred using 7 alternative genes, which were chosen semi-randomly with a diversity ~0.03 substitutions per site. The 4 phylogenetic trees are presented in Additional file 2. None of the trees match the expected phylogeny, which can be seen in Figure 4. The tree inferred from alternative genes and T. Whittam's scheme, seems to give the most discriminatory power.

Distribution of functional annotations

All genes were annotated with functional categories, where possible, using the COG database [13,14]. The annotations for the quarter of HGCs with the highest nucleotide diversity ("Most variable genes") and the quarter of HGCs with the lowest nucleotide diversity ("Most conserved genes") are compared in Figure 5.

Core-gene tree

The core-gene tree of *E. coli* is presented in Figure 6. A core-gene tree of the entire *Escherichia* genus is also presented as a small inset in Figure 6. The bootstrap values are scaled from 0 to 1, and indicate the fraction of the 500 bootstrap trees that agrees with each of the nodes. Bootstrap values of 1 are replaced with a black circle and bootstrap values between 0.7 and 1 are replaced by a grey circle. The tree containing all bootstrap values can be found in Additional file 3. The four main phylotypes A, B1, B2 and D are marked by the colors blue, red, purple and green, respectively. These phylotypes were determined *in silico*, based on the work done by Clermont *et al.* [15]. Additional phylotypes, C, E, and F, have also been reported [7,16,17] and are marked with their corresponding letters in Figure 6.

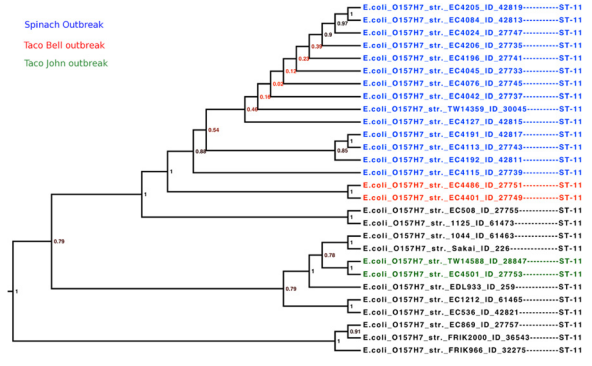

Figure 4 Core-gene tree close-up on O157:H7 strains. The tree is a close-up of the O157:H7 clade from the core-gene tree presented in Figure 6. The names has been colored according to the three outbreaks described in [21]. Blue strains represent the spinach outbreak, red strains represent the Taco Bell outbreak and the green strains represent the Taco John outbreak. Branch lengths have been modified to create the best visual output and thus have no value.

In 2009 Walk *et al.* [18] reported five novel phylogenetic clades, which were phylogenetically distinct from traditional *E. coli*, but they were unable to discriminate the novel clades from *E. coli* by traditional phenotypic profiling. These are sometimes referred to as Environmental *E. coli* or the cryptic *Escherichia* lineages. In 2011 Luo *et al.* sequenced strains from four of the five novel clades [19]. The four cryptic lineages are included in the Figure 6 inset and named Clade I, III, IV, and V. Clade I is included in the *E. coli* core tree as an outgroup because Clade I is very close to traditional *E. coli*. Clade I consists of 5 genomes, two of which have not, to our knowledge, been reported as Clade I strains. Using an *in silico* version of the identification procedure proposed by Clermont *et al.* [20], we further confirmed that the strains "*E. coli* STEC 7v" and "*E. coli* 1.2741" are indeed Clade I strains.

As a rule of thumb, bootstrap values above 0.7 are trustworthy, and in the core-gene tree in Figure 6, the bootstrap values are, in general, above this threshold.

Figure 4 presents a close-up of the ST 11 group of the core-gene tree. These results are in agreement with the SNP tree of a previous study on American O157:H7 outbreaks [21].

Pan-genome tree

The pan-genome tree is presented in Figure 7. The bootstrap values range from 0% to 100%, and indicate the percentage of the 500 bootstrap trees that agrees with each of the nodes. Bootstrap values of 100 are replaced with a black circle and bootstrap values between 70 and 100 are replaced with a grey circle. Bootstrap values below 70 are replaced with red circles. The tree containing all bootstrap values can be found in Additional file 4. The phylotypes are colored as in the core-gene tree (Figure 6).

Validation of methods

The standard deviation of all HGCs was calculated and plotted. The Alignments of the 10 HGCs with the highest standard deviation were examined and the gene sequences were BLASTed against the nr database, Uniprot, and annotated with protein domains using InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/). The HGCs seem to be well defined. The HGCs were either manually annotated as virulence factors (*e.g.* adhesins) or were of unknown function. Common to these 10 HGCs is also a very large average gene size. For the HGC with greatest standard deviation (adhesin) the average genes size is ~13,000 nucleotides. See Additional file 5 for details.

Genes were annotated with functional categories using the COG database. Each gene can be annotated with several categories. In this study it will be referred to as

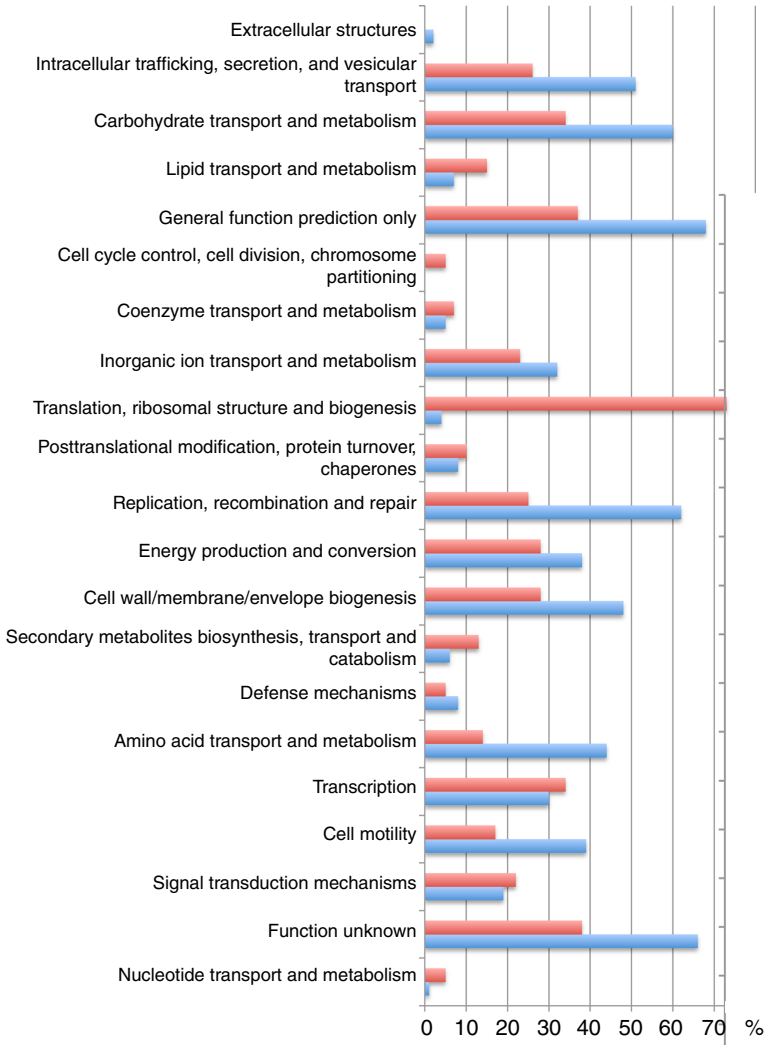

Figure 5 General function of conserved and variable HGCs. The difference in functional annotations between conserved and variable HGCs. Conserved here defined as the quarter of HGCs with the lowest nucleotide diversity (red bars) and variable defined as the quarter of HGCs with the highest nucleotide diversity (blue bars). Each HGC has a functional profile. A functional profile consists of one or more functional categories. The bars represent the percentage of HGC profiles, which contain the functional category listed to the immediate left of the bars.

the "functional profile'" of the gene. Ignoring the functional profile "unknown function", 4,123 HGCs contained genes with an identical profile. 12,189 HGCs could not be annotated. 59 HGCs contained genes with two different profiles, and 2 HGCs contained genes with more than two profiles. These two HGCs were examined and seem to be well defined. The 4,123 HGCs annotated with a single profile represents ~75% of all the genes.

In this study we include both draft and completed genomes. To estimate whether or not inclusion of draft sequences influences nucleotide diversity, we tested three datasets. One consisted of the 50 complete genomes, the other two consisted of 50 draft genomes randomly picked (without replacement). Clustering and nucleotide diversity calculation for all three datasets were performed. The two pan-genomes of the draft sequences seemed to be slightly higher than for the complete one. Virtually no difference in the distribution of nucleotide diversity was observed. See Additional file 6.

Discussion

In this study we identified core-genes and estimated the genetic variation among 186 publically available *E. coli* and *Shigella* genomes. Here, we will have a brief look at how *E. coli* is currently classified, how it fits our data, and discuss how these results may form a basis for future implementation of WGS as a standard typing tool for classification of *E. coli* in phylogeny and epidemiology and understanding *E. coli* evolution.

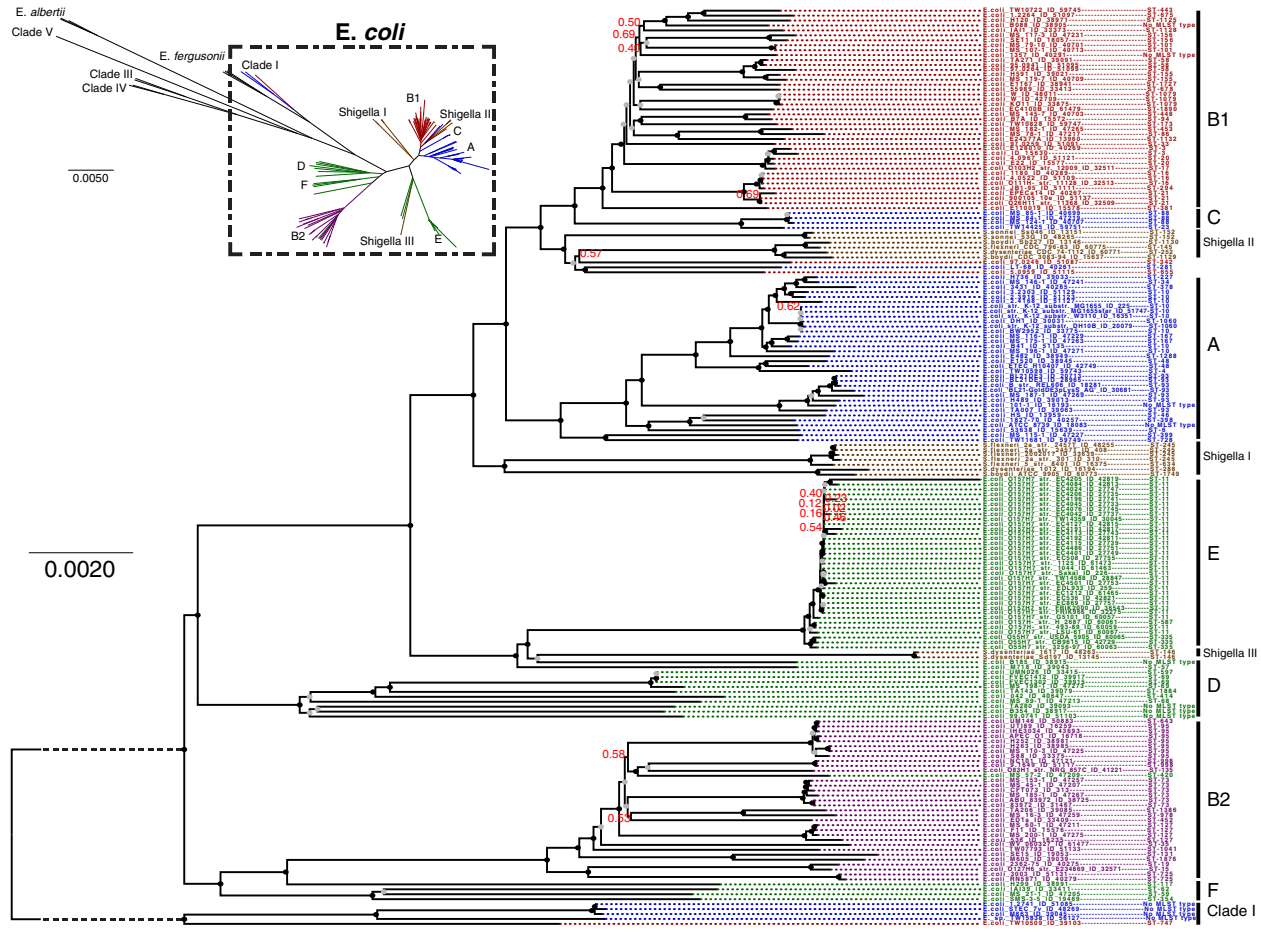

Figure 6 Core-gene tree. The *E. coli* tree was created from the alignment of 1,278 core-genes from the 186 *E. coli* genomes. MLST types are annotated to the far right of each genome name. The *Escherichia* genus tree was created from 297 core-genes. The phylotypes, as determined by the *in silico* Clermont [15] method, are marked with the colors blue (A), red (B1), purple (B2), green (D), and the *Shigella* genomes are marked with the color brown. At each node a black circle indicates a bootstrap value of 1, a grey circle a bootstrap value between 1 and 0.7 and a red number indicate an actual bootstrap value below 0.7. The dashed line in the figure represents a branch, which has been manually shortened by the authors to fit the figure on a printed page. The original tree with all bootstrap values can be seen in Additional file 2. Both trees are unrooted, but the *E. coli* tree has been visually rooted on the node leading to Clade I.

The dataset analyzed was obtained from GenBank and is publically available from NCBI. Two data quality issues are immediately encountered when using sequence data produced by others and from several different researchers: genome annotation and sequence quality. The annotation of the sequences can be very different, due to different annotation pipelines. Some annotations are manually curated and others are not. The completeness of each sequence can vary – some completed sequences are more "complete" than others. Chain *et al.* suggested a list of 6 categories in which all sequenced genomes could be defined based on their level of completeness [22]. In an attempt to overcome the bias from different annotations all genomes were annotated using the Prodigal gene finder [23] which provided consistency across the entire data set.

Sequence quality is also a concern. Unfortunately there hasn't been much focus on the issue, and publications estimating error rates in sequence databases are scarce. To our knowledge there are no recent publications estimating error rates in bacterial genomes deposited in GenBank. Wesche *et al.* estimated error rates in the mouse DNA sequences deposited to GenBank in 2004 [24]. They found an error rate of 0.1% in coding DNA sequences. This is lower than the estimate done in 1988 for all GenBank sequences deposited at the time, which demonstrated an error of ~0.3% [25].

Eukaryotes in general have much more complex genomes, due to introns, exons and complex repeats, which in turn leads to a higher than expected error rate. Sequencing technologies and assembly have also improved significantly since 1988. It is hypothesized that

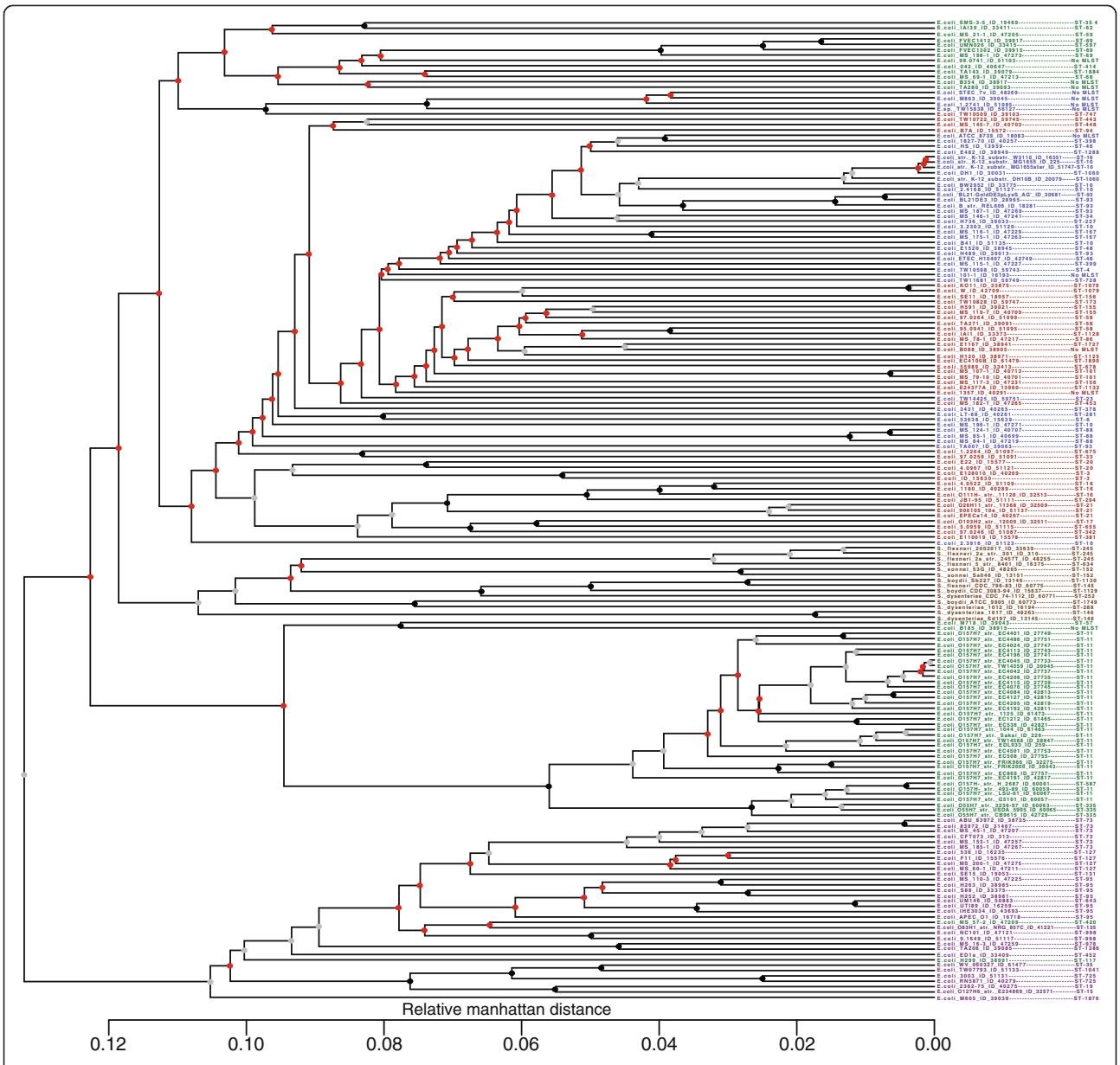

Figure 7 Pan genome tree. The tree was created based on the presence or absence of 16,373 HGCs in the 186 *E. coli* genomes. MLST types are annotated to the far right of each genome name. The phylotypes are marked with the colors blue (A), red (B1), purple (B2), green (D), and the *Shigella* genomes are marked with the color brown. Bootstrap values are annotated at each node as a percentage between 0 and 100. At each node a black circle indicates a bootstrap value of 100, a grey circle indicates a bootstrap value between 100 and 70 and a red circle indicates a bootstrap value below 70. The original tree with all bootstrap values can be seen in Additional file 3.

a conservative estimate of sequence errors in bacterial sequences deposited to GenBank today is less than 0.1%. Consequently an average *E. coli* gene (~1000bp) will contain approximately 1 error per gene.

Most errors caused by NGS technologies comes from insertions and deletions (indels), which will be completely ignored, due to the way nucleotide diversity is calculated. Therefore the errors, which are actually having an effect on the nucleotide diversity calculations, are probably lower than 0.1%. Because of these facts, it is believed that errors will, at most, cause 0.001 additional diversity to any of the variation calculations, and we believe that this is probably a very conservative estimate.

Sequencing errors, both indels and nucleotide changes can, however, cause genes to be truncated. Touchon *et al.* showed that at least 23 essential housekeeping genes were missing in their core-genome [7], and genomes missing these genes turned out to contain truncated versions of the "missing" genes. It was hypothesized that this was probably due to sequencing errors. Owing to the

possibility of sequencing errors accidently "deleting" genes from a genome, we also present the results for the soft core in this study.

Another issue, which sets a limit on our ability to interpret the results, is the lack of metadata, or specifically, the lack of a method for obtaining relevant metadata in an automated way. The amount of sequence data available now makes it unfeasible to email the corresponding author for each available genome to obtain its metadata. The community is aware of the increasing need for metadata and The Genomics Standards Consortium has suggested the Minimum Information about a Genome Sequence (MIGS), some of which is being incorporated into more recent GenBank files [26].

Pan- and core-genome

The core-genomes of *E. coli* and *Shigella* have been estimated in several studies. Lukjancenko *et al.* estimated the core-genome in 2010, from 61 genomes, using a single linkage clustering method and found it to be 1,472 HGCs if only *E. coli* was considered [5]. Vieira *et al.* estimated the core-genome in 2010 from 29 *E. coli* and *Shigella* genomes using the orthoMCL algorithm and found the core-genome to consist of 1,957 gene clusters [27]. In 2004 Fukiya *et al.* examined the core-genome from 22 *E. coli* strains using comparative genomic hybridization and estimated it to consist of approximately 2,800 shared open reading frames among all the strains [28]. Willenbrock *et al.* used high-density micro arrays to estimate the core-genome of 32 *E. coli* and *Shigella* genomes, and estimated the core-genome to be around 1,563 genes [29]. Chattopadhyaya *et al.* estimated the core-genome to consist of 1,513 genes among the 14 *E. coli* strains considered in their study [30]. Touchon *et al.* estimated the core-genome in 20 *E. coli* to be 1,976 genes and the pan-genome to consist of 11,432 genes. Thus, in previous studies (with fewer genomes) the size of the core-genome seems to fluctuate between 1,000 and 3,000 genes and generally conforms to the expectation that the core-genome would decrease, as an increased number of strains are analyzed, which might be an artifact of truncated genes due to sequencing errors.

In this study we found the soft core-genome to consist of 3,051 HGCs (Figure 1) for 186 genomes. In contrast to previous studies, we allowed a soft core-gene to be missing in up to 5% of all the genomes. If the strict core (HGC must be found in all genomes) was considered, the core-genome shrinks to 1,702 HGCs. It fits well within previous estimations made with the same strict cutoff.

The pan-genome has also been estimated in many studies and will probably continue to increase as more genomes are sequenced. In one study, the pan-genome of *E. coli* has been estimated to be as large as 45,000 gene families [31]. Another study suggests that the bacterial pan-genome is infinite [9]. Additional *E. coli* isolates, including some more distinctly related to those already sequenced, should be sequenced to obtain a more complete picture of the *E. coli* pan-genome.

Gene variation

The joint core-genome diversity plotted in Figure 2 (yellow) has one large top, which suggests that for most core-genes there is little room for diversity. Several smaller tops are also observed. We examined some HGCs that are part of the larger of the smaller tops (~0.17 substitutions per site). In both cases the HGC consisted of a gene coding for an enzyme and its isozyme counterpart. As for the case of one of the most diverse core families, the ABC transporters, the high diversity is due to different genes coding for proteins having very similar functions.

The pan-genome diversity plotted in Figure 2 has one large top and the distribution is much broader, as would be expected, due to the inclusion of the accessory genes.

No single, officially recognized system for classification of prokaryotes exists at the present time. The "polyphasic approach" is the most popular, and includes phenotypic, chemotaxonomic and genotypic data [32]. As for the genotypic data, this means that two genomes have to be 70% similar in order to be considered the same species. It has been shown that >70% similarity corresponds to an average nucleotide identity among the core-genes of >95% [32]. These results are supported by the median ~0.018 substitutions per site for the joint core found in this study.

Figure 3 shows that the genes from the Mark Achtman MLST scheme and the T. Whittam MLST scheme, in general, have less diversity than the majority of core HGCs. This is a bit surprising because the more variation in a gene, the greater the potential to be able to distinguish different strains.

The Pasteur MLST scheme seems to contain quite diverse core-genes, but also contains some which are more conserved than the average core-genes. This raises the question of whether or not a selection of more variable core-genes could be made, which, in turn, could provide higher resolution. Variability is, of course, not the only consideration when choosing MLST genes, *e.g.* an MLST scheme should not contain genes that are candidates for horizontal gene transfer, they should not be paralogous, and they should reflect the true phylogeny as much as possible. It is beyond the scope of this study to present a new MLST scheme, but it will be demonstrated how resolution could improve by choosing more diverse MLST genes. 7 core HGCs were chosen semi-randomly, with variation around ~0.03 substitutions per site. Genes were chosen with variation higher than average, although

not so high as to include paralogous genes. We found the corresponding genes in a set of 24 O157:H7 strains, aligned them and built a phylogenetic tree. Phylogenies were also inferred using each of the other three MLST schemes (see Additional file 2). We compared the MLST phylogenies with a published SNP tree created from these strains [21]. There is almost no variation found in the traditional Mark Achtman MLST scheme genes in these strains. In the alternate MLST scheme tree there is more variation and in turn more resolution. T. Whittam's scheme has the best overall resolution, probably due to the fact that T. Whittam's scheme contains twice as many genes as the other MLST schemes. None of the MLST phylogenies presents the expected topology. It seems unlikely that any selection of genes this small will ever be able to infer a robust phylogeny for an *E. coli* outbreak. At this point in time, there is probably no need to chase after a better MLST scheme, as WGS will probably make MLST typing obsolete with time. For most scientists, WGS is already less expensive than MLST typing [33]. WGS is, in general, far more promising, since it enables the use of entire core-genomes and SNPs (see core-gene tree discussion).

Barrick *et al.* [34] documented the mutations fixed in a specific *E. coli* strain over 40,000 generations *in vitro*. We looked at the genes and their corresponding HGCs in which these mutations occurred, but found no significant trend with regard to the variability of the mutated genes (data not shown).

Gene function distribution

Most HGCs could not be annotated with a functional category (~12,000); this corresponds to ~25% of all the genes.

The annotations of the HGCs are presented in Figure 5. As expected, the conserved genes are overrepresented in the "ribosomal" category, and even though there are only a few HGCs found in the "extracellular" category, they are exclusively from the variable HGC pool.

Core-gene tree

E. coli as a species contains within it a large diversity of adaptive paths. This is the result of a highly dynamic genome, with a constant and frequent flux of insertions and deletions [7,16]. Touchon *et al.* shows that the dynamic genome is compatible with a clonal population structure such as *E. coli*, since most gene acquisitions and losses happen in the exact same locations ("hotspots"). Hence the phylogenetic signal is still strong within the core genome even though recombination and lateral gene transfer is frequent [7].

The concatenated gene tree in Figure 6 demonstrates this strong phylogenetic signal quite well by the high fraction of confident nodes (confident nodes having a bootstrap value above 0.7). The tree also agrees with the MLST types. None of MLST types are actually split with the exception of ST-10, ST-11 and ST-93. In the ST-93 clade there is a single strain, which could not be typed by the *in silico* MLST algorithm. It is the draft genome of *E. coli* 101–1. Perfect matches for all 7 alleles are found, for the MLST scheme, but the combination is unknown. Its location within the ST-93 clade is valid though, since the unknown type is due to a single locus change (fumC-11 –> fumC-130). *E. coli* H 2687 with ST-587 is also a single locus variant of ST-11. ST-10 is split by ST-1060 and ST-167. Since the two strains of ST-1060 are sub-strains of K12, which is classified as ST-10, these fit inside the ST-10 clade. ST-167 is a single locus variant of ST-10.

All phylogroups (A, B1, B2, C, D, E, and F) also correspond very well with the core-gene tree. Only a few strains seem to violate the groups. *E. coli* MS 57 2 is classified as D, but the tree strongly suggests that it should belong to the B2 group. Gordon *et al.* showed that using the Clermont PCR multiplex method could lead to erroneous classification of phylotypes [35], in particular, classifying B2 phylotypes as D phylotypes were shown to be frequent. They proposed a new gene target, "ibeA", which will distinguish most B2 types from D types. *E. coli* MS 57 2 contains the gene target ibeA, which confirms its placement within the B2 phylogroup [35].

The tree supports the claim that B2 and F are the ancestral groups followed by D and then the sister groups B1 and A [7,16,36].

The fact that phylotyping and MLST typing fit so nicely with the core-gene tree, both confirms the highly clonal nature of *E. coli* and supports the use of core-genes to infer the "true" *E. coli* phylogeny.

To obtain a resolution high enough to be used in short term epidemiology, researchers have turned to inferring phylogenies from Single Nucleotide Polymorphism (SNP). SNP trees have, with much success, been used previously to describe complex outbreaks in detail [4,37]. However, to create a SNP tree, a good reference is needed and it is also frequently necessary to sort out false SNPs. The latter will always be subject to some controversy, because determination of a false SNP call will seldom be a completely objective call.

The creation of a core-gene tree requires no subjective alterations, which, in turn, also makes them much easier to automate and replicate than SNP trees. Figure 4 presents the E clade of the core-gene tree, and demonstrates the ability to differentiate three American *E. coli* O157:H7 outbreaks from each other. This is slightly better even, than the SNP tree published by Eppinger et. al [21].

In a case where the core-gene tree does not provide enough resolution, better resolution might be obtained

by focusing on the more variable genes; in these cases care should be taken not to focus on paralogous to infer phylogeny. Whether this is possible is doubtful, and will require further studies with strains of known origin and relationship for validation.

Based on many various typing methods, *Shigella* consistently has been shown to belong within the *E. coli* species [5]. Indeed, within Figure 6, all *Shigella* species can be seen to fall within the *E. coli* clade. How *Shigella* got the 'shiga toxin' and other pathogenicity genes has two opposing theories. One theory suggests that all the "*Shigella* genes" originated from one ancestral plasmid [38]. Another theory suggests that *Shigella* originated from three different *E. coli* species, which, independently of each other, acquired the "*Shigella* genes" [39]. Our core-gene tree (Figure 6) supports the latter theory, which is not surprising, since the theory was based on trees created from housekeeping genes. The core-gene tree fails to group the *Shigella* species. *Shigella* are classified based on their virulence factors, which are probably poor phylogenetic targets, and thus does not explain the "true" relationship between the *Shigella* species.

Pan-genome tree

The pan-genome tree is based on the absence or presence of all the HGCs of the pan-genome. It has been reported by Touchon *et al.* that gene conversion events are more likely than point mutations in *E. coli*. From this they conclude that the contribution made by recombination events outweigh site-level mutations as an evolutionary mechanism [7].

The pan-genome tree differs from the core-gene tree, because it is focused on those genes that are absent between the genomes. Since all the core-genes will be present in all genomes these will not in any way influence the phylogenetic relationship in this tree.

The pan-genome tree does not have as confident nodes as the core-gene tree. The deeper nodes are almost all below 50%. However, the nodes close to the leaves are quite confident and a majority of these reaches 70-100%.

These results are in agreement with the previously mentioned study by Touchon *et al.* The gene diversity in *E. coli* creates a poor phylogenetic signal between distantly related strains, since the signal is only made up from very few fixed ancestral insertions. This is due to the high gene flux in *E. coli* which causes only closely related strains to share a significant amount of accessory genes [7].

There are many similarities between the core-gene tree and the pan-genome tree, but also some obvious differences. The pan-genome tree does not divide the strains as nicely into the different phylogroups as the core-gene tree. The MLST type clades are also more divided than is the case for the core-gene tree. These results might not be that surprising, since both phylogroups and MLST types are based on a small set of core-genes and the pan-genome tree actually ignores these genes.

The pan-genome tree, due to one single *Shigella* clade, supports the "one origin" theory, as opposed to the core-gene tree, which supports the "three origins" theory of *Shigella*. Since the definition of *Shigella* is based upon a group of genes which gives it its pathogenic characteristics, it makes perfect sense that the pan-genome tree, which focuses on gene presence/absence, is able to isolate the *Shigella* genus into one single clade.

This convergence for *Shigella* has been observed previously by calculating the "metabolic distance" between *E. coli* strains. Vieira *et al.* suggests that this inconsistency between genetic distance and metabolic distance is proof that the *Shigella* metabolic networks have evolved quickly by genetic drift [27].

Both trees fail to divide the *Shigella* genus into any species clades, which further supports the argument that the taxonomy within *Shigella* might not be optimal.

Future perspectives

The core-gene tree in this study had a surprising capability to differentiate between closely related outbreak strains. However, more resolution might be needed to infer phylogenies or detect short-term outbreaks. In these cases, it might prove useful to put more weight on the variable regions of the genome. Further studies are needed to decide if this is a meaningful approach.

The results found in this study may lay ground for further studies into how we might create a standardized method for defining *E. coli* strains. To do this, studies are needed in which *E. coli* strains from different outbreaks and with different degrees of relatedness are sequenced and compared. Although "Single Nucleotide Polymorphism" (SNP) analysis was not done in this study, SNP potentially could be a powerful typing technique and will need to be included in future studies. This will, however, make more sense with a dataset that has been selected for this purpose.

It is becoming more and more apparent that a global epidemiological detection system is important, and for a global collaboration to be successful, standards are crucial.

Conclusions

Genes across different *E. coli* genomes are, in general, very well conserved. A pan-genome of 16,373 HGCs was found. A soft core-genome of 3,051 HGCs was found using a 95% cutoff, meaning that each HGC had to be found in 95% of the genomes to be considered a "core" HGC. With no genomes lacking HGC, we reached a core genome of 1,702 HGCs.

A pan-genome tree was created based on the absence or presence of genes. This method demonstrated the convergence of the *Shigella* lifestyle.

A core-gene tree was created based on the concatenated alignments of the core-genes. The core-gene tree was able to classify MLST types and phylotypes. We found that most genes used for MLST typing are less diverse than the majority of core-genes.

The core-gene tree showed a surprising capability of distinguishing a set of O157:H7 outbreak strains, and even seemed to do better than a SNP tree [21] created from the same strains. Future studies into a global standard for *E. coli* typing, should include a core-gene tree method, possibly combined with resolution improvement by focusing on variable genome regions, the latter is doubtful and remains to be tested.

The use of WGS will make it possible to eliminate, or at least reduce, the need for several typing steps used in traditional epidemiology. We are convinced that WGS is the optimal way forward in studying the phylogeny and epidemiology of *E. coli*.

Methods

All genomes analyzed were downloaded from GenBank at the National Center for Biotechnology Information (NCBI - http://www.ncbi.nlm.nih.gov/) on the 18th of April 2011. All draft and complete genomes were downloaded; a few were excluded due to content and quality. Draft genomes with fewer than 104,000 base pairs, and/or in more than 1,000 contigs were excluded. "*Shigella* sp. D9" with Genbank project ID 32507 was also excluded due to some very odd behavior in our analysis. We ended up with 171 *E. coli* and 15 *Shigella* genomes. The list of the 186 genomes can be found in Additional file 7. For each genome we predicted tRNAs with tRNAscan-SE version 1.23 [40] and rRNAs using rnammer [41] while gene prediction (excluding partial genes) was done using Prodigal version 2.6 [23]; *in silico* phylotyping was performed using in-house software, based on the presence or absence, determined by BLAST [42], of the two genes *chuA*, and *yjaA*, as well as the segment TspE4.C2 (unpublished), as proposed by Clermont *et al.* [15], and the MLST typing *in silico* was done using the MLST predictor at http://www.genomicepidemiology.org/ [33]. The same set of tools was also used for all the annotated genomes in GenBank in order to obtain consistency in the gene comparisons. The differences between the annotations made in this study and the annotated genomes are listed in Additional file 7.

Homolog gene clusters (HGCs)

Genes with similar sequences are likely to have similar functions and homologous gene clusters (HGCs) are generated by sequence similarity. In the ideal case, all occurrences of a specific gene from all the genomes will cluster exclusively into the same HGC. Using BLAT [43] all genes from all genomes were aligned against each other. The settings for BLAT were set to an E-value of at least 10^{-5}. The MCL software, based on the Markov Clustering Algorithm, developed by van Dongen [44] was then used to create the HGCs from the BLAT alignments.

This clustering approach has previously been applied to both Campylobacter [45] and *E. coli* [27]. The MCL software also does the clustering in orthoMCL software/web-service [46] (orthomcl.org).

Estimation of variation within HGCs

Multiple alignments were made for all HGCs using MUSCLE version 3.8.31 [47]. The multiple alignments were then used as input to VariScan version 2.0 [48], which calculated the nucleotide diversity based on the method suggested by Nei & Li [49]. At the gaps in the alignments, at least 10% of the members (or at least 2) had to have non-gap characters in the gap position to be included in the diversity calculation of the alignment. The "member cut-off" parameter was also set to 50% and 90%, we detected virtually no difference in the diversity distributions (data not shown).

Core- and pan-genome

The core- and pan-genomes were defined by HGCs. The soft core-genome was defined as all HGCs that had members in at least 95% of the 186 genomes, equivalent to at least 177 genomes of the 186 genomes. The strict core-genome was defined as all HGCs that have members in all genomes. The pan-genome was defined as all HGCs.

Functional annotation

All genes were blasted against the COG database [13], hits with an E-value > 10^{-5} were considered significant; only the best hits (highest bit score) were extracted. The functional profile of the best hit was then assigned to the query gene.

HGCs were annotated with the functional profile, which was dominant between the members of the HGC. This also included "not in COG".

Core-gene tree

A core-gene tree was created for all the members of the *Escherichia* genus and another one was made for only *E. coli* and *Shigella*. Both are presented in Figure 6.

To create a core-gene tree, all genes not found in all genomes were removed. A multiple alignment for each gene was then done using MUSCLE version 3.8.31 [47]. The alignments were then concatenated. 500 resamples of the alignment were created with Seqboot version 3.67 [50]. Distance matrices were calculated for the initial alignment as well as for each of the 500 resamples using dnadist version 3.67 [50]. Trees were then created using FastME from NCBI [51] and the tree from the original

alignment was compared to the 500 trees from the resamples using CompareToBootstrap [52].

FigTree (http://tree.bio.ed.ac.uk/software/figtree/) has been used to visualize the final core-gene tree. The tree is unrooted, but has been visually rerooted with FigTree on the node leading to Clade I.

Pan-genome tree

A phylogenetic tree was created based upon the absence or presence of all HGCs and a hierarchical clustering based on calculations of the Manhattan distance between each HGC. Singletons were ignored. The tree was created with the R package, as previously described by Snipen & Ussery [53].

Additional files

Additional file 1: Genes used in MLST schemes. Lists of the three groups of genes used in the Mark Achtman, Pasteur institute, and T. Whittam MLST schemes.

Additional file 2: MLST phylogenies of O157:H7. Four phylogenetic trees inferred from four different MLST schemes. Tree A is inferred from Mark Achtman's MLST scheme, tree B is inferred from the Pasteur MLST scheme, tree C is inferred from T. Whittam's MLST scheme and tree D is inferred from the alternative MLST scheme used in this proof of concept case.

Additional file 3: Core tree with all bootstrap values. The tree was created from the alignment of each of the 1,278 core genes from the 186 *E. coli* genomes. MLST types are annotated to the far right of each genome name. The phylotypes are marked with the colors blue (A), red (B1), purple (B2), green (D), and the *Shigella* genomes are marked with the color brown.

Additional file 4: Pan-genome tree with all bootstrap values. The tree was created based on the presence or absence of 16,373 HGCs in the 186 *E. coli* genomes. MLST types are annotated to the far right of each genome name. The phylotypes are marked with the colors blue (A), red (B1), purple (B2), green (D), and the *Shigella* genomes are marked with the color brown. Bootstrap values are annotated at each node as a percentage between 0 and 100.

Additional file 5: Annotation of highly deviating HGCs. Manual annotation of the 10 HGCs with the highest standard deviation in gene size. The annotation is based on blasting the gene members against the nr database, Uniprot and running the sequences through InterProtScan.

Additional file 6: Complete versus draft nucleotide diversity distributions. The nucleotide diversity distribution is plotted for both the core-HGCs and the pan-HGCs of the three datasets: complete (red), draft1 (blue), and draft2 (green).

Additional file 7: Table of complete dataset. The table shows the dataset used for the article. The "GB genes" column indicates the number of genes annotated in the corresponding GenBank file. The "Prod genes" column indicates the number of genes that was found with prodigal for this study.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RSK carried out comparative genomics analysis, did interpretation of results and drafted the manuscript. CF helped do the comparative genomics analysis, did interpretation of the results and helped to draft the manuscript. DWU and FMA helped with interpretation of the results and to draft the manuscript. All authors were involved in conception and design. All authors have read and approved the manuscript.

Acknowledgements

This study was supported by the Center for Genomic Epidemiology at the Technical University of Denmark and funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

Author details

[1]DTU Food, The Technical University of Denmark, Kgs Lyngby, Denmark. [2]Department of Systems Biology, Center for Biological Sequence Analysis, The Technical University of Denmark, Kgs Lyngby, Denmark.

Received: 10 September 2012 Accepted: 22 October 2012
Published: 31 October 2012

References

1. Otto TD: **Real-time sequencing.** *Nat Rev Microbiol* 2011, **9:**633–633.
2. Oryan M, Prado V, Pickering L: **A millennium update on pediatric diarrheal illness in the developing world.** *Semin Pediatr Infect Dis* 2005, **16:**125–136.
3. Russo TA, Johnson JR: **Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem.** *Microbes Infect* 2003, **5:**449–456.
4. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *N Engl J Med* 2011, **364:**730–739.
5. Lukjancenko O, Wassenaar TM, Ussery DW: **Comparison of 61 sequenced escherichia coli genomes.** *Microb Ecol* 2010, **60:**708–720.
6. Bromham L, Penny D: **The modern molecular clock.** *Nat Rev Genet* 2003, **4:**216–224.
7. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EPC, Denamur E: **Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths.** *PLoS Genet* 2009, **5:**e1000344.
8. Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proc Natl Acad Sci U S A* 1979, **76:**5269–5273.
9. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome.** *Trends Genet* 2009, **25:**107–110.
10. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M: **Sex and virulence in Escherichia coli: an evolutionary perspective.** *Mol Microbiol* 2006, **60:**1136–1151.
11. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E, Lortholary O, Clermont O, Denamur E, Picard B, Nassif X, Brisse S: **Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains.** *BMC Genomics* 2008, **9:**560.
12. Qi W, Lacher D, Bumbaugh A, Hyma K, Quellette L, Large T, Whittam: **EcMLST: an online database for multi locus sequence typing of pathogenic escherichia coli.** *Comput Syst Bioinformatics Conf* 2004, 520–521.
13. Tatusov RL: **A genomic perspective on protein families.** *Science* 1997, **278:**631–637.
14. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinforma* 2003, **4:**41.
15. Clermont O, Bonacorsi S, Bingen E, Bonacorsi P: **Rapid and Simple Determination of the Escherichia coli Phylogenetic Group.** *Appl Environ Microbiol* 2000, **66:**4555–4558.
16. Tenaillon O, Skurnik D, Picard B, Denamur E: **The population genetics of commensal Escherichia coli.** *Nat Rev Microbiol* 2010, **8:**207–17.
17. Escobar-Páramo P, Clermont O, Blanc-Potard A-B, Bui H, Le Bouguénec C, Denamur E: **A specific genetic background is required for acquisition and expression of virulence factors in Escherichia coli.** *Mol Biol Evol* 2004, **21:**1085–94.
18. Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam. TS: **Cryptic lineages of the genus Escherichia.** *Appl Environ Microbiol* 2009, **75:**6534–44.

19. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT: **Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species.** *Proc Natl Acad Sci USA* 2011, **108:**7200–5.
20. Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E: **Characterization of the cryptic Escherichia lineages: rapid identification and prevalence.** *Environ Microbiol* 2011, **13:**2468–77.
21. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA: **Genomic anatomy of Escherichia coli O157: H7 outbreaks.** *Proc Natl Acad Sci* 2011, **108:**20142–20147.
22. Chain P, Grafham D, Fulton R, Fitzgerald M, Hostetler J, Muzny D, Ali J, Birren B, Bruce D, Buhay C, *et al*: **Genome project standards in a new era of sequencing.** *Science* 2009, **326:**236.
23. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics* 2010, **11:**119.
24. Wesche PL, Gaffney DJ, Keightley PD: **DNA sequence error rates in Genbank records estimated using the mouse genome as a reference.** *DNA Seq* 2004, **15:**362–364.
25. Krawetz SA: **Sequence errors described in GenBank: a means to determine the accuracy of DNA sequence interpretation.** *Nucleic Acids Res* 1989, **17:**3951–3957.
26. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, Vos PD, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Leebens-mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone S, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B: **The minimum information about a genome sequence (MIGS) specification.** *Nat Biotechnol* 2008, **26:**541–547.
27. Vieira G, Sabarly V, Bourguignon P-Y, Durot M, Le Fèvre F, Mornico D, Vallenet D, Bouvet O, Denamur E, Schachter V, Médigue C: **Core and panmetabolism in Escherichia coli.** *J Bacteriol* 2011, **193:**1461–72.
28. Fukiya S, Mizoguchi H, Tobe T: **Extensive genomic diversity in pathogenic Escherichia coli and Shigella strains revealed by comparative genomic hybridization microarray.** *J Bacteriol* 2004, **186:**3911–3921.
29. Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW: **Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray.** *Genome Biol* 2007, **8:**R267.
30. Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV: **High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection.** *Proc Natl Acad Sci USA* 2009, **106:**12412–7.
31. Snipen L, Almøy T, Ussery DW: **Microbial comparative pan-genomics using binomial mixture models.** *BMC Genomics* 2009, **10:**385.
32. Schleifer KH: **Classification of Bacteria and Archaea: past, present and future.** *Syst Appl Microbiol* 2009, **32:**533–42.
33. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Pontén TS, Ussery DW, Aarestrup FM, Lund O: **Multilocus Sequence Typing of Total Genome Sequenced Bacteria.** *J Clin Microbiol* 2012, **33:**1355–1361.
34. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF: **Genome evolution and adaptation in a long-term experiment with Escherichia coli.** *Nature* 2009, **461:**1243–7.
35. Gordon DM, Clermont O, Tolley H, Denamur E: **Assigning Escherichia coli strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method.** *Environ Microbiol* 2008, **10:**2484–96.
36. Sims GE, Kim S: **Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs).** *PNAS* 2011, **108:**8329–8334.
37. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science (New York, N.Y.)* 2010, **327:**469–474.
38. Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E: **The evolutionary history of Shigella and enteroinvasive Escherichia coli revised.** *J Mol Evol* 2003, **57:**140–8.
39. Pupo GM, Lan R, Reeves PR: **Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics.** *Proc Natl Acad Sci USA* 2000, **97:**10567–72.
40. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955–64.
41. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes.** *Nucleic Acids Res* 2007, **35:**3100–8.
42. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389–3402.
43. Kent WJ: **BLAT--The BLAST-Like Alignment Tool.** *Genome Res* 2002, **12:**656–664.
44. Dongen S: *Graph Clustering by Flow Simulation.*: Proefschrift Universiteit Utrecht; 2000.
45. Lefébure T, Bitar PDP, Suzuki H, Stanhope MJ: **Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept.** *Genome Biol Evol* 2010, **2:**646–55.
46. Li L, Jr CS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178–2189.
47. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792–7.
48. Hutter S, Vilella AJ, Rozas J: **Genome-wide DNA polymorphism analyses using VariScan.** *BMC Bioinformatics* 2006, **7:**409.
49. Nei M, Li W: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proc Natl Acad Sci U S A* 1979, **76:**5269–5273.
50. Felsenstein J: **PHYLIP - Phylogeny Inference Package.** *Cladistics* 1989, **5:**164–166.
51. Desper R, Gascuel O: **Fast and accurate phylogeny minimum-evolution principle.** *J Comput Biol* 2002, **9:**687–705.
52. Price MN, Dehal PS, Arkin AP: **FastTree 2–approximately maximum-likelihood trees for large alignments.** *PLoS One* 2010, **5:**e9490.
53. Snipen L, Ussery DW: **Standard operating procedure for computing pangenome trees.** *Stand Genomic Sci* 2010, **2:**135–141.

doi:10.1186/1471-2164-13-577
Cite this article as: Kaas *et al.*: Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics* 2012 **13**:577.

BMC Genomics

PROCEEDINGS Open Access

snpTree - a web-server to identify and construct SNP trees from whole genome sequence data

Pimlapas Leekitcharoenphon[1,2*], Rolf S Kaas[1,2], Martin Christen Frølund Thomsen[2], Carsten Friis[1], Simon Rasmussen[2], Frank M Aarestrup[1]

From Asia Pacific Bioinformatics Network (APBioNet) Eleventh International Conference on Bioinformatics (InCoB2012)
Bangkok, Thailand. 3-5 October 2012

Abstract

Background: The advances and decreasing economical cost of whole genome sequencing (WGS), will soon make this technology available for routine infectious disease epidemiology. In epidemiological studies, outbreak isolates have very little diversity and require extensive genomic analysis to differentiate and classify isolates. One of the successfully and broadly used methods is analysis of single nucletide polymorphisms (SNPs). Currently, there are different tools and methods to identify SNPs including various options and cut-off values. Furthermore, all current methods require bioinformatic skills. Thus, we lack a standard and simple automatic tool to determine SNPs and construct phylogenetic tree from WGS data.

Results: Here we introduce snpTree, a server for online-automatic SNPs analysis. This tool is composed of different SNPs analysis suites, perl and python scripts. snpTree can identify SNPs and construct phylogenetic trees from WGS as well as from assembled genomes or contigs. WGS data in fastq format are aligned to reference genomes by BWA while contigs in fasta format are processed by Nucmer. SNPs are concatenated based on position on reference genome and a tree is constructed from concatenated SNPs using FastTree and a perl script. The online server was implemented by HTML, Java and python script.
The server was evaluated using four published bacterial WGS data sets (*V. cholerae*, *S. aureus* CC398, *S.* Typhimurium and *M. tuberculosis*). The evalution results for the first three cases was consistent and concordant for both raw reads and assembled genomes. In the latter case the original publication involved extensive filtering of SNPs, which could not be repeated using snpTree.

Conclusions: The snpTree server is an easy to use option for rapid standardised and automatic SNP analysis in epidemiological studies also for users with limited bioinformatic experience. The web server is freely accessible at http://www.cbs.dtu.dk/services/snpTree-1.0/.

Background

The dramatic decrease in cost for whole-genome sequencing (WGS) has made this technology economically feasible as a routine tool for scientific research, including infectious disease epidemiology. In addition, WGS has major applications for health service providers working with infectious diseases [1] as such to deliver high-resolution genomic epidemiology as the ultimate typing method for bacteria.

The ideal microbial typing technique should enable differentiation of epidemiological unrelated strains and group epidemiological related (outbreak) strains, [2] and give information that will help to understand the evolutionary history of multiple strains within a clonal lineage [1,2]. Although some current technologies are highly informative like MLST or PFGE, they have limited resolution when applied to closely related isolates and different methods often have to be applied in different situations [1,2].

* Correspondence: pile@food.dtu.dk
[1]National Food Institute, Building 204, Technical University of Denmark, 2800 Kgs Lyngby, Denmark 4444
Full list of author information is available at the end of the article

Especially outbreak isolates normally have very little diversity and require extensive genomic methods to differentiate and catagorize the isolates [3]. Single nucleotide polymorphisms (SNPs) also show relatively low mutation rates and are evolutionarily stable. Moreover, SNPs analysis has successfully been used for determining broad patterns of evolution in many recent studies [4-6].

Currently, There are a number of available non-commercial NGS genotype analysis software such as SOAP2 [7], GATK [8] and SAMtools [9]. Nonetheless, all of the software require bioinformatic skills, various options, various setting and they do not have a user friendly web-interface.

Here we introduce snpTree. A server for online-automatic SNP analysis and SNP tree construction from sequencing reads as well as from assembled genomes or contigs. The server is a pipeline which intregrates avaliable SNPs analysis softwares such as SAMtools [9] and MUMmer [10], with customized scripts. The performance of the server was evaluated with four published bacterial WGS data set; *Vibrio cholerae* [3], *Staphylococcus aureus* CC398 [6], *Salmonella* Typhimurium [11] and *Mycobacterium tuberculosis* [12].

Implementation

The snpTree server was created to handle both WGS data and assembled genomes to generate a phylogenetic tree based on SNPs data. The overall process is shown in Figure 1. For raw reads (Figure 1A), snpTree use an in-house toolbox (Genobox) for mapping and genotyping which consists of avaliable programs for next-generation sequencing analysis such as Burrows-Wheeler Aligner, BWA [13] and software package for SNPs calling and genotyping, SAMtools [9]. The source code of Genebox is available at https://github.com/srcbs/GenoBox. For contigs or assembled genomes (Figure 1B), MUMmer [10] is used for both reference genome alignment and SNPs identification processes.

The web-server contains more than 2,000 completed reference genomes collected from NCBI Genome database (accessed on April 2012).

SNPs identification from WGS

Prior to mapping raw reads to a proper reference genome, the sequence data in fastq format are filtered and trimmed according to the following criteria [14]: (i) reads with N's are removed, (ii) if a read matches a minimum of 25 nt of a sequencing primer/adaptor the reads are trimmed at the 5' coordinate of match, (iii) the 3' tail bases are trimmed if the quality score is less than 20, (iv) the minimum average quality of the read should be 20 and the read length after trimming should be at least 20 nt.

Trimmed raw reads are aligned against a reference genome using BWA [13] with minimum mapping quality equal to 30 as a default (Figure 1A). BWA is based on an effective data compression algorithm called Burrows-Wheeler transform (BWT) that is fast, memory-efficient and espectially useful for aligning short reads [15].

SNPs calling and filtering are accomplished by SAMtools that is a software package for parsing and manipulating alignments in the generic alignment format (SAM/BAM format) [9]. The snpTree server allows users to set a couple of parameters to filter SNPs, a minimum coverage and a minimum distance between each SNPs (prune). The default for both cut-offs is set to 10 and additionally all heterozygous SNPs are filtered because these are likely mapping errors in haploid chromosomes. The identifed SNPs are concluded into a VCF file.

SNPs identification from assembled genomes

A pipeline has been developed around the software package MUMmer version 3.23 [10] (Figure 1B). An application named Nucmer, which is part of MUMmer, is used to align each of *de novo* assemblies to a reference genome chosen by the user (default settings). SNPs are then called from the resulting alignments with another MUMmer application named "show-snps" (with options "-CIlrT"). A pruning is then applied, if chosen by the user, and the SNPs are written into a VCF formatted file for each of the analyzed genomes.

SNPs tree construction

One VCF formatted file is needed for each Operational Taxonomic Unit (OTU). The SNPs are then concatenated into a single alignment by ignoring indels. Including indels would disturb the position of SNPs in the sigle alignment. To include indels in any trees, it requires some sensible way to represent them numerically as distances in an evolutionary space, and there is no any ways to achieve this. Indels could theoretically be included in a multiple sequence alignment, since such alignments can handle gaps but it's difficult to score them. "Blast-like" gap penalties certainly would not work, since they are optimized for much larger gaps, e.g. recombination events.

It is important to note that SNPs not found in a VCF file is interpreted as not being a variation and the corresponding base in the reference is expected. This might not always be the right choice, because a SNP not found in a VCF file could be a result of an INDEL. It is expected to be a rare case and probably won't disturb the phylogenetic signal.

The alignment is passed on to Fastree [16], which creates a maximum likelihood tree from the SNP alignment.

snpTree server output

snpTree server provides an output to users with SNPs tree figure in SVG format, number of SNPs and other relevant output files such as (i) SNPs files, which contains

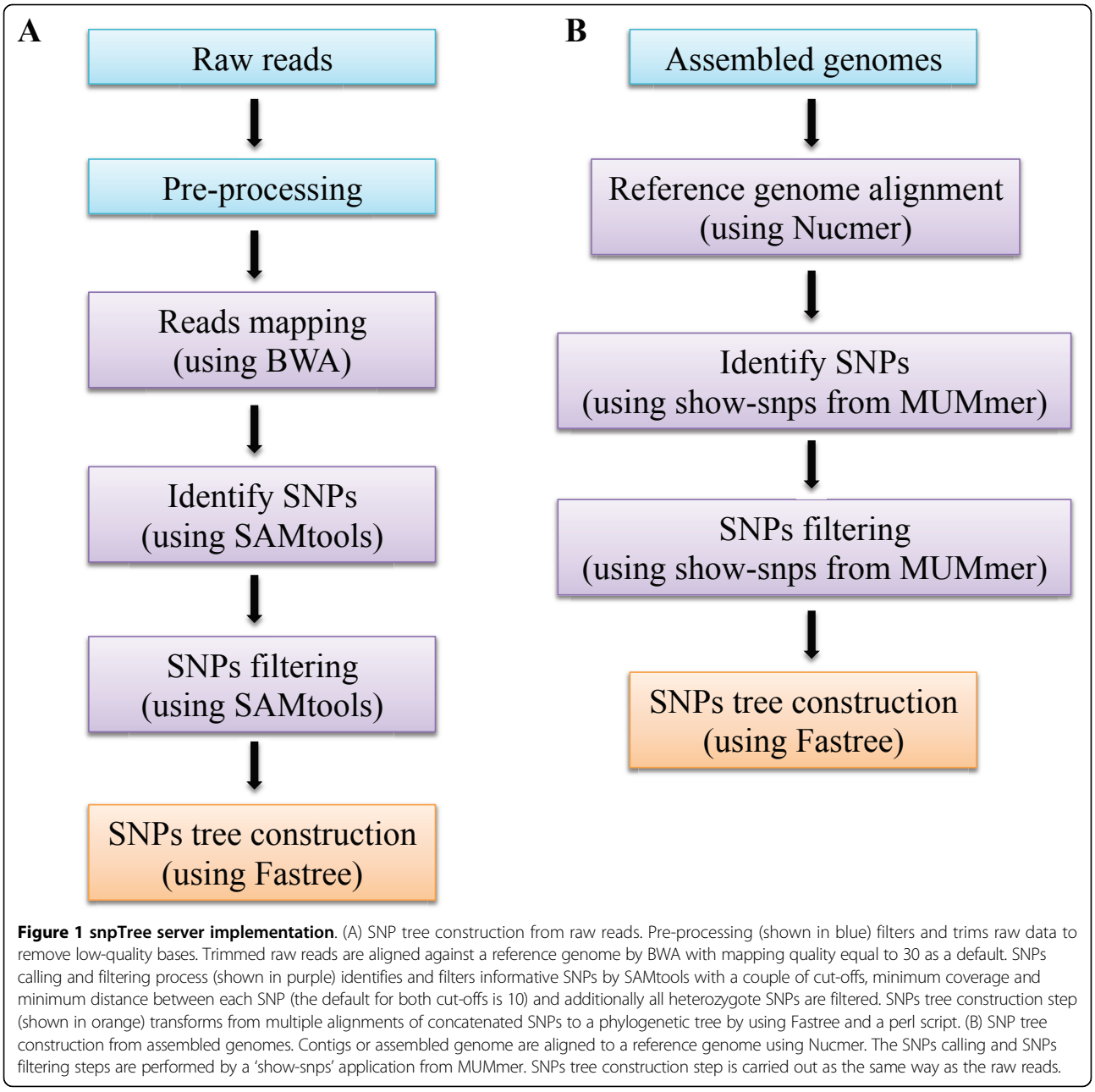

Figure 1 snpTree server implementation. (A) SNP tree construction from raw reads. Pre-processing (shown in blue) filters and trims raw data to remove low-quality bases. Trimmed raw reads are aligned against a reference genome by BWA with mapping quality equal to 30 as a default. SNPs calling and filtering process (shown in purple) identifies and filters informative SNPs by SAMtools with a couple of cut-offs, minimum coverage and minimum distance between each SNP (the default for both cut-offs is 10) and additionally all heterozygote SNPs are filtered. SNPs tree construction step (shown in orange) transforms from multiple alignments of concatenated SNPs to a phylogenetic tree by using Fastree and a perl script. (B) SNP tree construction from assembled genomes. Contigs or assembled genome are aligned to a reference genome using Nucmer. The SNPs calling and SNPs filtering steps are performed by a 'show-snps' application from MUMmer. SNPs tree construction step is carried out as the same way as the raw reads.

identified SNPs including indels for each input genome in VCF format [17], (ii) cancatenated SNPs in newick, phylip and fasta format, (iii) SNPs annotation files which give users an overview of nucleotide changes or amino acid changes from SNPs including which input genomes contain which SNPs as well as information about synonymous and non-synonymous SNPs (Additional file 1). An example of output is shown in Figure 2.

Results and discussion

The snpTree was evaluated using raw reads and assembled genomes from four published bacterial WGS data sets (*V. cholerae* [3], *S. aureus* CC398 [6], *S.* Typhimurium [11] and *M. tuberculosis* [12]). The evaluation was considered based on tree topology as well as the reference genome's position of identifed SNPs.

Evaluation of tree topology and SNPs position

WGS from published data set were subjected to snpTree server in order to generate SNP trees. The tree topology evaluation was based on percentage of concordance. If the strain in the tree from snpTree server matches exactly with the tree from published data, it was considered as an exact match. If the strains were grouped into

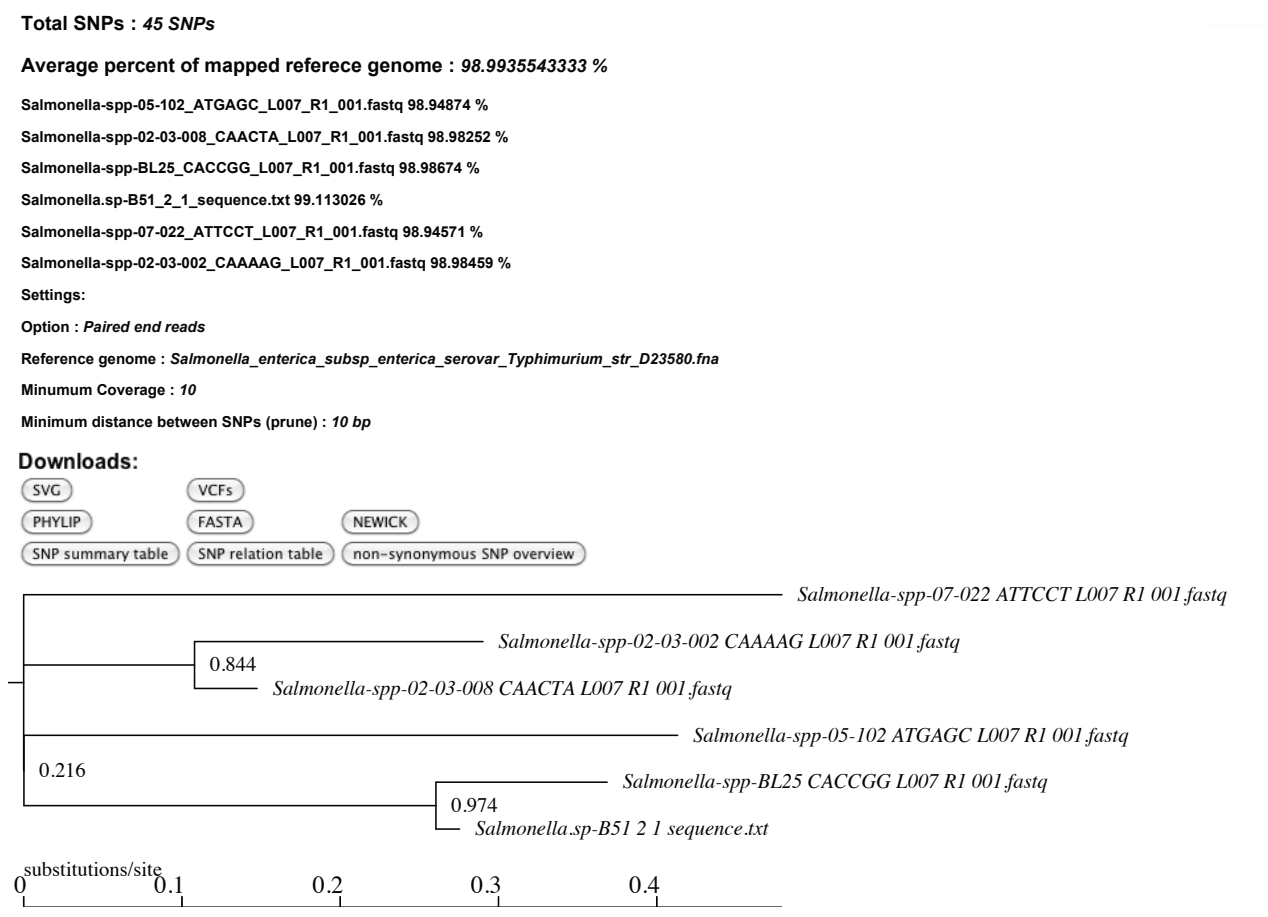

Figure 2 snpTree output. An example of the output from snpTree server using Illumina paired-end reads as input data.

the same cluster with published data, it was considered as a cluster match. In addition, the snpTree server was evaluated with assembled genomes or contigs. The raw reads were assembled prior by *de novo* assembly using Velvet 1.1.04 [18]. The assembled genomes were processed to snpTree server to make SNP trees.

V. cholerae data set

The evaluation results are summarized in Table 1. For the *V. cholerae* data set, the performance of snpTree from raw reads (Figure 3) and contigs (Additional file 2) were accurate in term of exact match and cluster match. From Figure 3, all of genomes were grouped into the same clusters as in the original tree. In the Nepal-1 cluster, there are only 3 genomes that are not in the same position compared to the original tree. However, the isolates in Nepal-1 group are highly homogeneous and there are some synapomorphic SNPs (genome position that has mutated the new nucleotide which shared with all descendants) supporting its unique identities [3].

The percentage of overlapped and non-overlapped SNPs between published data and snpTree server is illustrated in Figure 4A for raw reads and Figure 4B for assembled genomes. For *V. cholera*, both raw reads and contigs (Figure 4), the snpTree server identified SNPs mostly from the same position in published data (95%

Table 1 Evaluation table

Data set	Percentage of concordance	
	Exact match	cluster match
V. cholerae (raw reads)	91	100
***V. cholerae* (contigs)**	**85**	**100**
S. aureus CC398 (raw reads)	88	96
***S. aureus* CC398 (contigs)**	**87**	**97**
S. typhimurium (raw reads)	61	100
***S. typhimurium* (contigs)**	**53**	**100**
M. tuberculosis (raw reads)	58	78
***M. tuberculosis* (contigs)**	**25**	**72**

The percentage of concordance from comparing SNP trees from snpTree server against the four published data set.

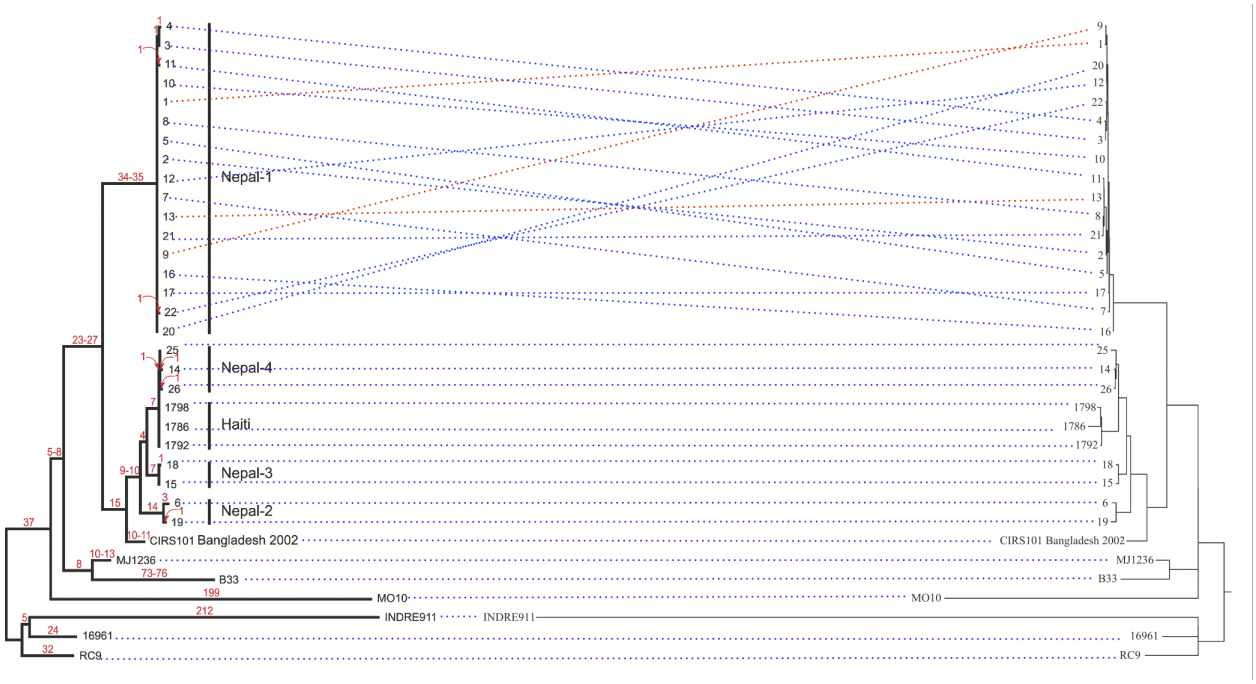

Figure 3 Comparison between phylogenetic trees from published data set (*V. cholerae*) and snpTree server. These trees (34 WGS from ***V. cholerae***) shows comparison of tree topology between the trees from original publication (left) and snpTree server (right). The linked lines indicate exact match for each genome in the tree. According to the tree from published data, the blue lines mean exact match and the red one represent inexact match.

overlapped SNPs). This result supports the consistency of the tree from snpTree server (Figure 3).

S.aureus CC398 data set

For *S. aureus* CC398 (Table 1), snpTree produced a tree with 87 - 88 % concordance for exact match and 96 - 97 % concordance for cluster match. SNP trees for raw reads and assembled genomes are shown in Additional file 3 and Additional file 4 respectively. There were 91 and 90 % overlapping SNPs for raw reads and assembled genomes (Figure 4). The performance of snpTree on this data set was slightly less than for the *V. cholera* data set. The reason is probably that the genomes of 89 *S. aureus* CC398 isolates came from animals and humans sources from 19 countries and four continents. In addition, there are 4,238 SNPs among them [6]. These isolates are more diverse than *V. cholera* isolates. Thus, this diversity makes difficulty for snpTree to capture exactly the same variant as in original publication. Nevertheless, snpTree can differentiate between isolates from humans and pigs which is very meaningful to epidemiological studies.

S. Typhimurium data set

The third data set, *S.* Typhimurium, which consists of 51 *Salmonella* in which 43 isolates from 14 patients with multiple recurrences in Blantyre, Malawi and 8 control typhimurium isolates [11]. Like in the original publication, both raw reads and contigs data set, the isolates fell within three distint phylogenetic clusters (Additional file 5 and 6) which gave 100 % concordance for cluster match (Table 1). On the other hand, the percentage of concordance for exact match was quite low (53 - 61 %). It is not possible to evaluate SNPs position for this data set because of lacking SNPs position data. However, the number of identified SNPs from snpTree server (1,692 SNPs) was not much different from original data set (1,463 SNPs). Most of the *S.* Typhimurium isolates are highly genetically related as they came from patients who had recrudescence and/or reinfections. Therefore, this study requires high-resolution SNPs analysis and intensive phylogenetic tree construction to differentiate these little variation. In addition, the original tree from this data set was generated and confirmed using several independent approaches, with bootstrap support and clade credibility marked [11] which snpTree cannot repeat as using bootstrapping is time-consuming.

M.tuberculosis data set

Another data set that consists of 32 *M. tuberculosis* outbreak isolates and 4 historical isolates (from the same region but isolated before the outbreak) with matching genotype suggesting that the outbreak was clonal [12].

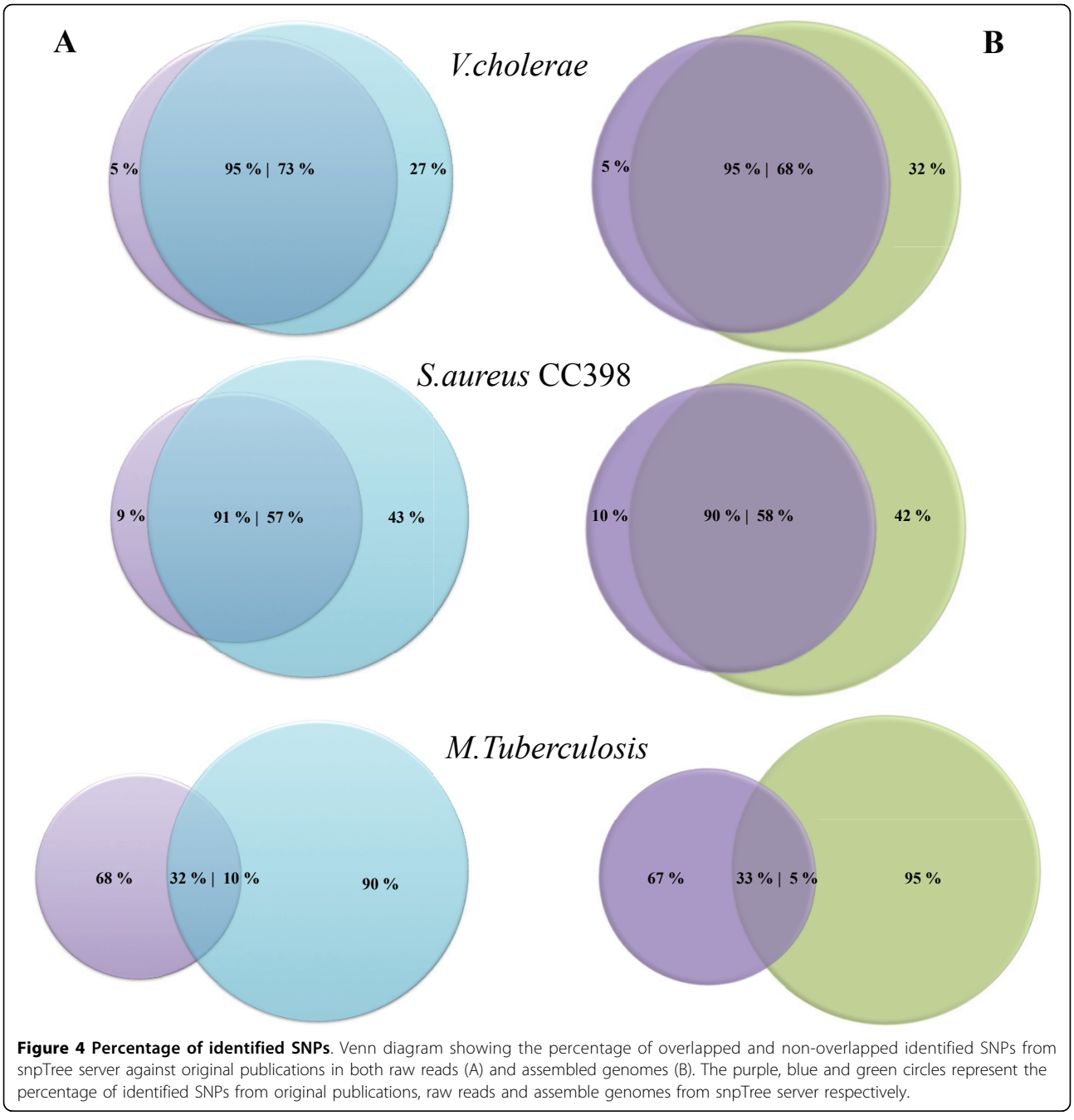

Figure 4 Percentage of identified SNPs. Venn diagram showing the percentage of overlapped and non-overlapped identified SNPs from snpTree server against original publications in both raw reads (A) and assembled genomes (B). The purple, blue and green circles represent the percentage of identified SNPs from original publications, raw reads and assemble genomes from snpTree server respectively.

The performance of snpTree server on this data set was inconsistent due to low concordance percentage for exact match and cluster match (Table 1, Additional file 7 and 8). Moreover, the number of indentified SNPs and matching SNP positions (Figure 3) are very different between the tree from snpTree server (677 SNPs) and the published data (204 SNPs). The original publication determined transmission dynamics of the outbreak at a higher resolution by filtering to remove many of SNPs in repetitive regions and those appearing in a single isolate. Thus, the procedure in the original manuscript is impossible to repeat and it should be noted that the original filtering reduced the number of SNP's from more than 1,000 to 204. This is probably the reason that snpTree were unable to reproduce the same results as in the original publication.

Sensitivity and specificity

In order to evaluate the sensitivity and specificity of SNP calling method, the artificial sequence was created

from a genome of 4,878,012 bp with 1,000 randomly SNP artificial inserted. The simulated sequence was aligned to a reference genome and identified SNPs using SNP idenfication pipeline for assemble genome. SNPs calling was performed with varied two cut-off values which are minimum number of bp between SNPs (prune) and minimum number of bp from a sequence end (e). The sensitivity and specificity for SNP identification were summarized in Table 2.

The sensitivity for prune cut-off (Table 2) was slightly dropped when increasing number of prune. This is due to the more number of bp between SNPs (prune) leading to the high chance to have SNPs between that number of bp.

Using minimum number of bp from a sequence end as a varied cut-off, the sensitivity was very high and stable for all varied values. It is quite rare to have SNPs occurred in the tails of sequence so this cut-off less affects to the SNP calling process. The specificity for both cut-off were very high. It is because the number of SNP inserted is extreamly low (1,000 SNPs) compared to the whole genome (4,878,012 bp).

The rapid technological advantages in WGS and rapidly decreasing cost has made the technology available for large groups of scientists as well as clinical microbiologists. It is expected that WGS will very soon find widespread use in clinical and public health microbiology, as has already been shown [19]. The implementation of such technologies will however, create a major need for simple to use bioinformatic tools to make sense of the data generated. We have here developed snpTree and evaluated it on four different published datasets. The concordance of the SNPs tree from raw reads was more adequate than the one from assembled genomes, which is not surprising. However, in practice transfering sequencing reads will be more time-consuming than just transferring assembled genomes and the tree topology from these different kind of genomes was only sligthly different. Therefore, the assembled genomes option in snpTree server can provide a quicker solution for uploading time-consuming. In order to create informative SNPs tree, using a closely related reference genome is important. Therefore, the selection of a proper reference genome is crucial. Thus, it is adviced to choose a reference genome belonging to the same or as closely related a sub-type as possible to the strain collection under study. This could for species where this is a available reference belonging to the same MLST type. In the future a more generic solution to overcome this obstracle might be to using high-resolution prediction method such as K-mers to assign a genuine reference genome.

Table 2 Sensitivity and specificity

Variable and cut-off value	Sensitivity (%)	Specificity (%)
Number of bp between SNPs		
0	97.8	100
10	97.2	99.99988
25	96.6	99.99975
50	95.8	99.99959
75	94.6	99.99935
100	93.8	99.99918
Number of bp from a sequence end		
0	97.8	100
10	97.8	100
25	97.8	100
50	97.8	100
75	97.8	100
100	97.7	100

Evaluation of sensitivity (SN) and specificity (SP) using different settings of minimum number of bp between SNPs (prune) and minimum number of bp from a sequence end (e) for SNP detection on a simulated dataset consisting of a genome of 4,878,012 bp with 1,000 randomly SNP artificial inserted.

Conclusions

The advance of WGS and the use of epidemiological genomics underline the potential of practical application of WGS for clinical microbiology and emphazies the importance of biology and evolution in developing reliable and accurate genomics tools for clinical use. In addition, SNP-typing phylogenetic methods can distinguish very closely related isolates to a degree not achievable by widely employed sub-genomic typing tools. snpTree server might be not a perfect tool but it is an option for easy and rapid standardised and automatic SNP analysis tool in epidemiological studies. It is also useful for users with limited bioinformatic experience.

Additional material

Additional file 1: Example of SNP annotation output.

Additional file 2: SNP trees from contigs of *V. cholerae* data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 3: SNP trees from raw reads of *S. aureus* CC398 data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 4: SNP trees from contigs of *S. aureus* CC398 data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 5: SNP trees from raw reads of *S.* Typhimurium data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 6: SNP trees from contigs of *S.* Typhimurium data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 7: SNP trees from raw reads of *M. tuberculosis* data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 8: SNP trees from contigs of *M. tuberculosis* data set (left is the tree from original publication and right is the tree from snpTree server).

Acknowledgements
This study was supported by the Center for Genomic Epidemiology (09-067103/DSF) http://www.genomicepidemiology.org and Danish Food Industry Agency (3304-FVFP-08). PL and RKM would like to acknowledge funding from the Technical University of Denmark.
This article has been published as part of *BMC Genomics* Volume 13 Supplement 7, 2012: Eleventh International Conference on Bioinformatics (InCoB2012): Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S7.

Author details
[1]National Food Institute, Building 204, Technical University of Denmark, 2800 Kgs Lyngby, Denmark 4444. [2]Center for Biological Sequence Analysis, Building 208, Department of Systems Biology, Technical University of Denmark, 2800 Kgs Lyngby, Denmark.

Authors' contributions
PL planned the study, carried out web-server construction and drafted the manuscript. RKM constructed SNPs analysis pipeline for assembled genomes and automatic SNP tree construction pipeline. MCFT participated in web-server construction. CF constructed automatic SNPs tree construction pipeline. SR constructed SNPs analysis pipeline for raw reads and developed Genobox toolbox. FMA supervised, planned the study and drafted the manuscript. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 13 December 2012

References

1. Parkhill J, Wren BW: **Bacterial epidemiology and biology–lessons from genome sequencing.** *Genome biology* 2011, **12**:230.
2. Foxman B, Zhang L, Koopman JS, Manning SD, Marrs CF: **Choosing an appropriate bacterial typing technique for epidemiologic studies.** *Epidemiologic perspectives & innovations* 2005, **2**:10.
3. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM: **Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak.** *MBio* 2011, **2.**
4. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327**:469-74.
5. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Møller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP: **Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:3065-70.
6. Price LB, Stegger M, Hasman H, Aziz M, Larsen J, Andersen PS, Pearson T, Waters AE, Foster JT, Schupp J, Gillece J, Driebe E, Liu CM, Springer B, Zdovc I, Battisti A, Franco A, Zmudzki J, Schwarz S, Butaye P, Jouy E, Pomba C, Porrero MC, Ruimy R, Smith TC, Robinson DA, Weese JS, Arriola CS, Yu F, Laurent F, Keim P, Skov R AF: **Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock.** *MBio* 2012, **3**:1-6.
7. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J: **SNP detection for massively parallel whole-genome resequencing.** *Genome Res* 2009, **19**:1124-1132.
8. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome research* 2010, **20**:1297-303.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-9.
10. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic acids research* 2002, **30**:2478-83.
11. Okoro CK, Kingsley RA, Quail MA, Kankwatira AM, Feasey NA, Parkhill J, Dougan G, Gordon MA: **High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal salmonella typhimurium disease.** *Clinical infectious diseases* 2012, **54**:955-63.
12. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *The New England journal of medicine* 2011, **364**:730-9.
13. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-60.
14. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O: **Multilocus Sequence Typing of Total Genome Sequenced Bacteria.** *Journal of clinical microbiology* 2012, 1355-1361.
15. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nature reviews. Genetics* 2011, **12**:443-51.
16. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.** *Molecular biology and evolution* 2009, **26**:1641-50.
17. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**:2156-8.
18. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome research* 2008, **18**:821-9.
19. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW: **A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance.** *BMJ Open* 2012, **2.**

doi:10.1186/1471-2164-13-S7-S6
Cite this article as: Leekitcharoenphon *et al.*: snpTree - a web-server to identify and construct SNP trees from whole genome sequence data. *BMC Genomics* 2012 **13**(Suppl 7):S6.

OPEN ACCESS Freely available online

PLOS ONE

Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms

Rolf S. Kaas[1]*, Pimlapas Leekitcharoenphon[1], Frank M. Aarestrup[1], Ole Lund[2]

1 National Food Institute, Technical University of Denmark, Lyngby, Denmark, **2** Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark

Abstract

Whole genome sequencing (WGS) shows great potential for real-time monitoring and identification of infectious disease outbreaks. However, rapid and reliable comparison of data generated in multiple laboratories and using multiple technologies is essential. So far studies have focused on using one technology because each technology has a systematic bias making integration of data generated from different platforms difficult. We developed two different procedures for identifying variable sites and inferring phylogenies in WGS data across multiple platforms. The methods were evaluated on three bacterial data sets and sequenced on three different platforms (Illumina, 454, Ion Torrent). We show that the methods are able to overcome the systematic biases caused by the sequencers and infer the expected phylogenies. It is concluded that the cause of the success of these new procedures is due to a validation of all informative sites that are included in the analysis. The procedures are available as web tools.

Citation: Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O (2014) Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. PLoS ONE 9(8): e104984. doi:10.1371/journal.pone.0104984

Editor: Alex Friedrich, University Medical Center Groningen, Netherlands

Received March 19, 2014; **Accepted** July 14, 2014; **Published** August 11, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All 80 files are available from the ENA/SRA/DDBJ databases. Accession numbers: ERR277214 ERR493470 ERR493462 ERR493461 ERR493457 ERR277216 ERR493471 ERR493465 ERR493475 ERR493449 ERR277217 ERR493481 ERR493452 ERR493456 ERR493478 ERR493459 ERR493473 ERR493479 ERR493477 ERR493468 SRR353666 SRR446761 SRR500737 SRR359788 SRR446821 SRR500745 SRR354026 SRR446808 SRR500738 SRR354027 SRR446809 SRR500739 SRR359773 SRR446740 SRR500741 SRR354019 SRR446757 SRR500724 SRR354039 SRR446811 SRR500740 SRR364526 SRR446822 SRR500746 SRR364527 SRR446756 SRR500747 SRR364528 SRR446823 SRR500748 SRR359777 SRR446818 SRR500742 SRR354021 SRR446760 SRR500725 SRR445275 ERR493446 ERR493474 ERR493448 ERR493466 SRR445237 ERR493450 ERR493454 ERR493455 ERR493469 SRR445277 ERR493476 ERR493451 ERR493464 ERR493447 SRR445072 ERR493482 ERR493460 ERR493480 ERR493458 ERR493467 ERR493453 ERR493472 ERR493463.

Funding: This study was supported by the Center for Genomic Epidemiology (www.genomicepidemiology.org) grant 09-067103/DSF from the Danish Council for Strategic Research. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: rkmo@food.dtu.dk

Introduction

Microbial whole-genome sequencing using bench-top sequencing technologies holds great promises to enhance diagnostic and public health microbiology [1–3]. Its great value in describing and improving our understanding of bacterial evolution, outbreaks and transmission events has been shown in a number of recent studies, including *Staphylococcus aureus* [4–6], *Vibrio cholera* [7], *Escherichia coli* [8], *Mycobacterium tuberculosis* [9] and surveillance of antimicrobial resistance [10]. All of these studies have however, been done retrospectively, (except [8], which was done prospectively) and conducted using the same technology and performed in a single laboratory.

For rapid detection of out-breaks involving multiple sites or even countries it is essential to enable rapid and reliable comparison of data generated in different laboratories and using different technologies [1]. Enabling comparison between technologies is also important for the future comparison of data generated using novel technologies that are currently under development and comparison to data already generated using current technologies. An important step to enable this is to allow for sequencing platform independent analysis. This is especially relevant for SNP calling where the currently available sequencing platforms all have some type of systematic sequencing bias [11–16]. These systematic biases' today make it virtually impossible to perform reliable phylogenetic studies if the data are generated using different technologies. For research purposes the correct identification of SNP's might be solved by sequencing using multiple platforms, but for infectious disease out-breaks this will neither be practical or timely feasible. Infectious disease out-breaks are often multistate and rapid comparison and correct clustering is essential.

Common practice in SNP calling is to use a closely related reference genome, often a reference genome that has been sequenced and finished with respect to the study in question. While this approach is feasible for research purposes it is not practical in an out-break investigation.

In this study we developed two novel procedures for identifying variations in whole genome sequencing reads and conducting phylogenetic analysis of isolates. The procedures were evaluated on an available data-set where three different platforms had been used to sequence the same 12 *Salmonella* Montevideo isolates, as well as sequencing of selected *Salmonella* Typhimurium and *Staphylococcus aureus* isolates using Illumina and Life Technologies.

The novel procedures have been made available as web tools at the following addresses:

Nucleotide Difference (ND) method: http://cge.cbs.dtu.dk/services/NDtree/.

Table 1. Reference Genomes.

Ref. genome	Distance	Size (bp)	Accession No.
S. aureus CC398	close	2,872,582	AM990992.1
S. aureus ST228	distant	2,759,835	NC_020533.1
S. DT104	close	4,933,631	HF937208.1
S. Schwarzengrund	distant	4,709,075	NC_011094.1

doi:10.1371/journal.pone.0104984.t001

Novel SNP procedure: http://cge.cbs.dtu.dk/services/CSIPhylogeny/.

Materials and Methods

Datasets

Three different datasets were used for evaluation in the present study, comprising selected *Salmonella* Montevideo [17], *Staphylococcus aureus* CC398 [5], and *Salmonella* Typhimurium DT104 [18] from previous studies.

For *S.* Montevideo 12 closely related outbreak strains where sequenced once by US Food and Drug Administration using Roche Genome sequencer FLX system, Illumina MiSeq and Life Technologies Ion Torrent and made publicly available (Table S1), although only the MiSeq data was used in the original study [16]. The raw data were downloaded from the Sequence Read Archive (SRA). For *Staphylococcus aureus* CC398, the completely sequenced and annotated strain SO385 (AM990992.1) as well as four additional strains were selected from a previously published study [5] and sequenced twice using both MiSeq and Ion Torrent. HiSeq was used in the original study for sequencing. All the strains except for the reference strain were chosen from the same clade, named IIa1i in the original study. The strains are not epidemiologically related but have all been isolated from Danish Pigs and are shown to be closely related in the original study. For S. Typhimurium DT104 the reference strain NCTC 13348 (HF937208.1) and an additional three isolates from the same outbreak [18] were sequenced twice on both MiSeq and Ion Torrent.

Genomic DNA (gDNA) was purified from the isolates using the Easy-DNA extraction kit (Invitrogen) and DNA concentrations determined using the Qubit dsDNA BR Assay Kit (Invitrogen). The isolates were sequenced twice on the MiSeq platform (Illumina) and Ion Torrent PGM (Life Technologies).

For Ion Torrent the isolates were sequenced following the manufacturer's protocols for 200 bp gDNA fragment library preparation (Ion Xpress Plus gDNA and Amplicon Library 96 Preparation), template preparation (Ion OneTouch System), and sequencing (Ion PGM 200 Sequencing kit) using the 316 chip. For MiSeq the isolates chromosomal DNA of the isolates was used to create genomic libraries using the Nextera XT DNA sample preparation kit (Illumina, cat. No. FC-131-1024) and sequenced using v2, 2×250 bp chemistry on the Illumina MiSeq platform (Illumina, Inc., San Diego, CA).

Data analysis

The raw data was trimmed and cleaned for adapters using AdapterRemoval v. 1.1 (https://code.google.com/p/adapterremoval/) before any analysis was done.

The data were analyzed using an available and published pipeline for SNP-calling and creation of phylogenetic trees [19], a recently developed method based on nucleotide differences [18], as well as a novel procedure for SNP-calling developed in this study. All three methods requires a reference sequence, these has been listed in Table 1. All the references applied in this study are available as complete assemblies from GenBank.

Nucleotide Difference (ND) procedure (Novel). A previously published procedure [18] was used. In Brief, each read were mapped to the reference genome. A base was called if $Z = (X-Y)/\mathrm{sqrt}(X+Y)$ was greater than 1.96 corresponding to a p-value of 0.05. Here X is the number of reads X having the most common nucleotide at that position, and Y the number of reads supporting other nucleotides. It was further required that $X>10*Y$. The number of nucleotide differences in positions called in all sequences was counted, and a matrix with these counts was given as input to an UPGMA algorithm implemented in the neighbor program v. 3.69 (http://evolution.genetics.washington.edu/phylip.html) in order to construct the tree.

SNP analysis (Novel). Reads were mapped to reference sequences using BWA v. 0.7.2 [20]. The depth at each mapped position was calculated using genomeCoverageBed, which is part of BEDTools v. 2.16.2 [21]. Single nucleotide polymorphisms (SNPs) were called using mpileup part of SAMTools v. 0.1.18 [22]. SNPs were filtered out if the depth at the SNP position was not at least 10x or at least 10% of the average depth for the particular genome mapping. The reason for applying a relative depth filter is to set different thresholds for sequencing runs that yield very different amounts of output data (total bases sequenced). SNPs were filtered out if the mapping quality was below 25 or the SNP quality was below 30. The quality scores were calculated by BWA and SAMTools, respectively. The scores are phred-based but can be converted to probabilistic scores, with the formula $10\hat{}(-Q/10)$, where Q is the respective quality score. The probabilistic scores will represent the probability of a wrong alignment or an incorrect SNP call, respectively. In each mapping, SNPs were filtered out if they were called within the vicinity of 10 bp of another SNP (pruning). A Z-score was calculated for each SNP as described above for NDtree.

The depth requirements ensure that all positions considered are covered by a minimum amount of reads. The SNP quality and the Z-score requirements ensures that all positions considered are also called with significant confidence with respect to the bases called at each position.

All genome mappings were then compared and all positions where SNPs was called in at least one mapping were validated in all mappings. The validation includes both the depth check and the Z-score check as for the SNP filtering. Any position that fails validation is ignored in all mappings.

Maximum Likelihood trees were created using FastTree [23].

snpTree. Analysis was done using the method described by Leekitcharoenphon et al. [19]. The primary difference between the snpTree method and the novel SNP analysis is in the filtering and validation of the SNP positions. Briefly, the snpTree method calls SNPs using BWA [20], then the default behavior is to filter

Table 2. Comparison of the novel SNP procedure, the Nucleotide Difference (ND) method and snpTree.

Method	Percent of reference genome covered				
	S. Montevideo	*S.* DT104		*S. aureus*	
	Distant ref.	*Distant ref.*	*Close ref.*	*Distant ref.*	*Close ref.*
snpTree	100.00	100.00	100.00	100.00	100.00
novel SNP	81.40	92.48	99.42	93.05	99.40
ND	34.48	88.60	95.68	63.44	88.00
	Informative sites				
snpTree	22068	26691	79	20324	699
novel SNP	18 (36)	49	66	107	252
ND	19 (33)	54	66	126	602
	Average distance within clusters				
snpTree	6353.0	8024.0	8.1	4271.0	69.0
novel SNP	0 (0)	1.0	1.5	1.0	2.0
ND	0 (0)	0.0	0.0	2.0	3.6

doi:10.1371/journal.pone.0104984.t002

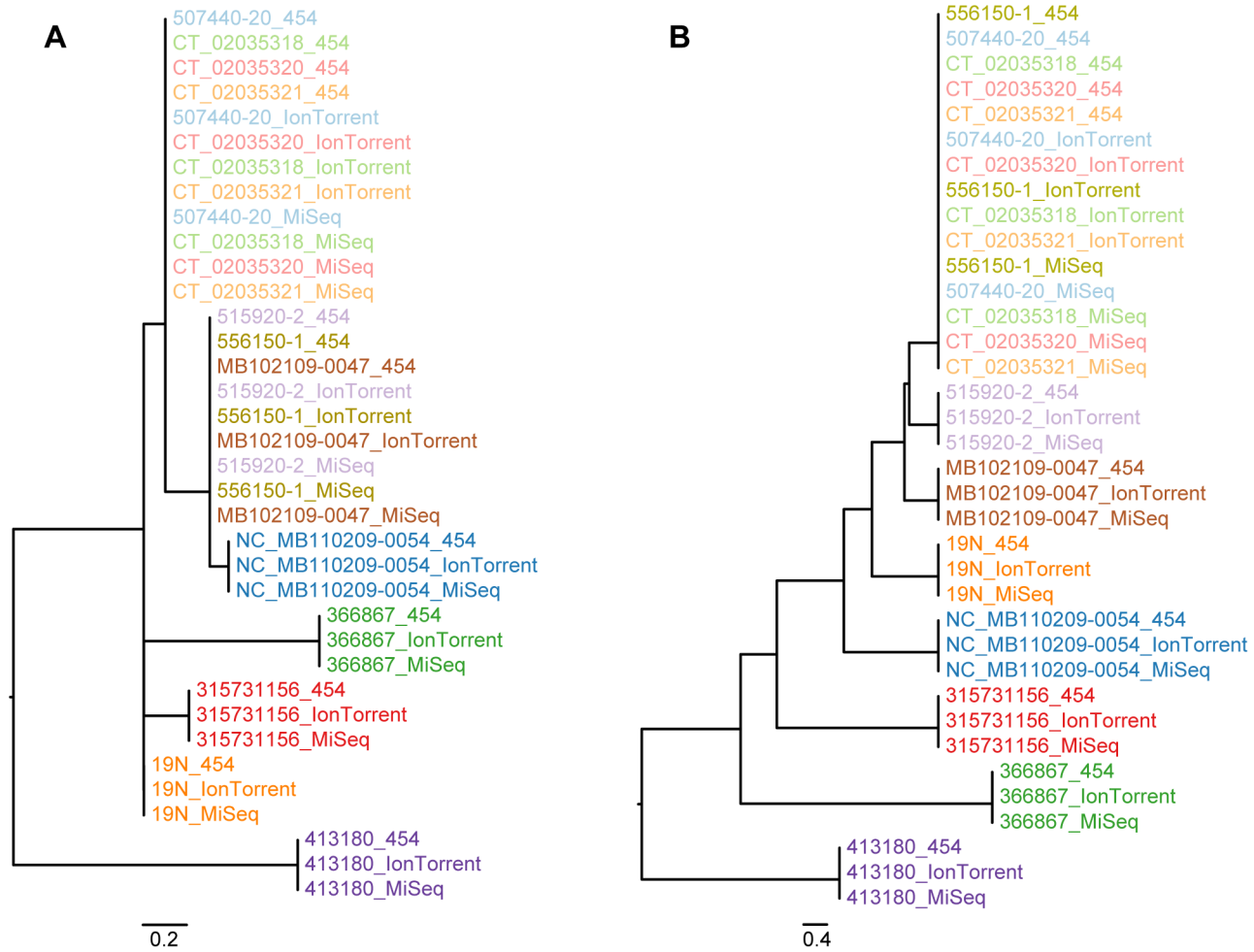

Figure 1. ***Salmonella*** **Montevideo phylogeny (complete dataset).** Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. (**A**) Phylogeny inferred with novel SNP procedure; (**B**) Phylogeny inferred with the Nucleotide Difference (ND) method.
doi:10.1371/journal.pone.0104984.g001

out SNPs with a depth less than 10 and SNPs found within 10 bps of each other (pruning). An alignment of the SNPs are then created by concatenating the SNPs. Positions where no SNPs are found or where SNPs has been ignored are assumed to be identical to the base in the reference sequence. A maximum likelihood tree is created from the alignment.

Results

A comparison of the three different methods is given in Table 2 and Figure 1–4 for the different datasets. The original procedure (snpTree) was un-able to cluster the same isolates correctly across the different technologies whereas both of the novel methods gave improved results.

snpTree does not ignore any positions and is potentially able to consider 100% of the genome. The novel SNP procedure considers between 81.40% and 99.42%. The ND method is more conservative and considers between 34.48% and 95.68%. The snpTree method was expected to have issues with the references that were distantly related as also mentioned by the authors of this method. This is also illustrated in Table 2 by the large amount of informative SNPs that the method finds compared to the other methods, when the references are distantly related to the analyzed isolates. A plot of the number of positions that each isolate causes to be ignored in the Montevideo analysis (see Figure S1) shows very clearly that three isolates causes more than half of the ignored positions. The three isolates were deemed of low quality, removed from the analysis, and the methods were rerun. The numbers from the rerun is presented in parentheses in Table 2.

Salmonella Montevideo

Each of the three methods was applied to just the MiSeq data and compared to the SNP tree published by Allard et al. [17] (Figures S2, S3, and S4). The novel SNP procedure infers a phylogeny that agrees with the published one. The ND procedure infers a tree that almost agrees with the published one, except that the "clinical clade" is reversed with respect to the most recent common forefathers. The snpTree method infers a phylogeny that is very different from the published one and will therefore not be discussed here (Figure S2).

Figure 1A and 1B presents the phylogeny that was inferred by applying the entire Montevideo dataset to the novel SNP procedure and the ND method, respectively. Compared to the MiSeq only phylogeny it is observed that the phylogeny has lost a lot of resolution, but in general keeps the same topology, as the respective phylogenies inferred with the MiSeq data alone.

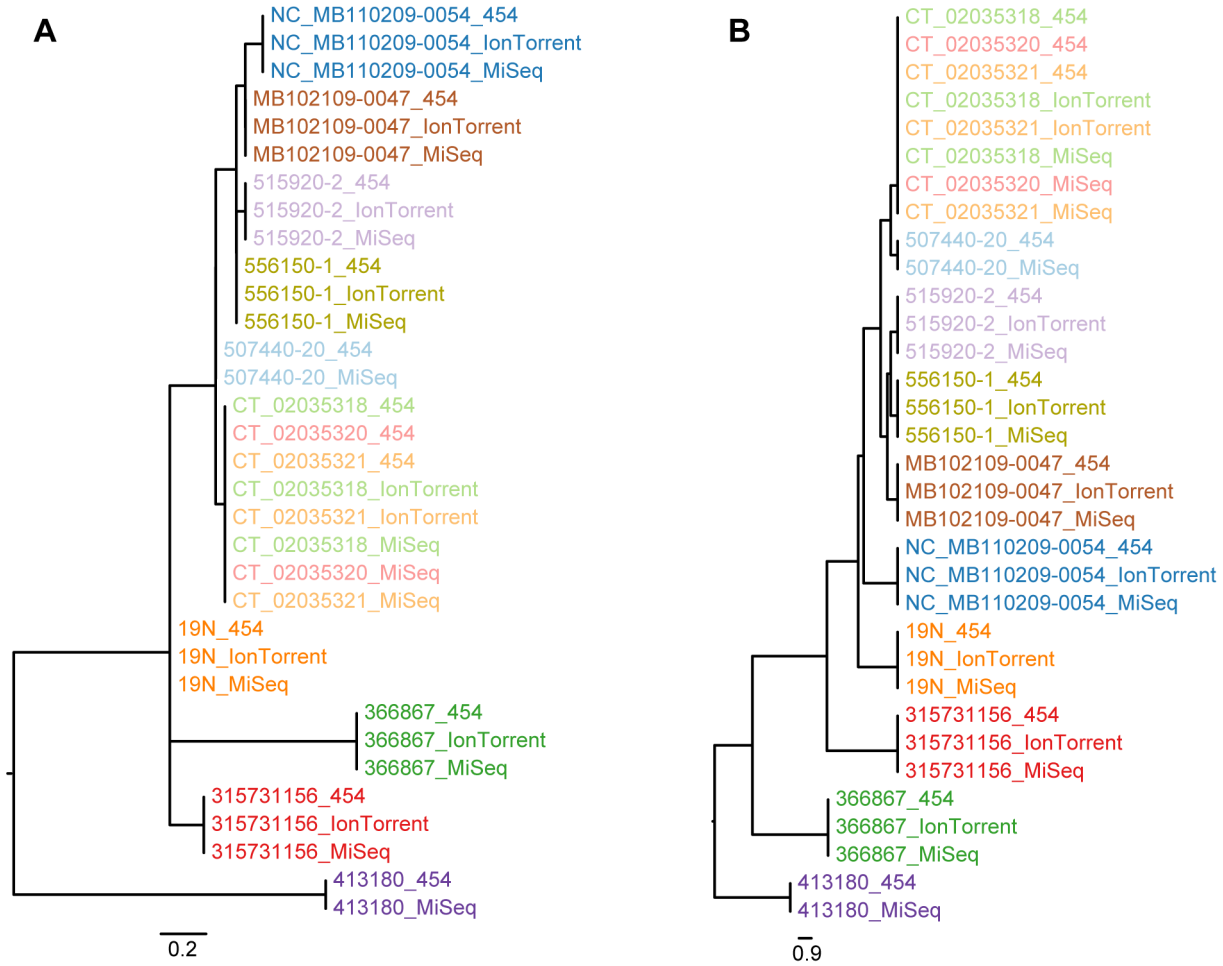

Figure 2. ***Salmonella*** **Montevideo phylogeny (low quality sequences removed).** Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. (**A**) Phylogeny inferred with novel SNP procedure; (**B**) Phylogeny inferred with the Nucleotide Difference (ND) method.
doi:10.1371/journal.pone.0104984.g002

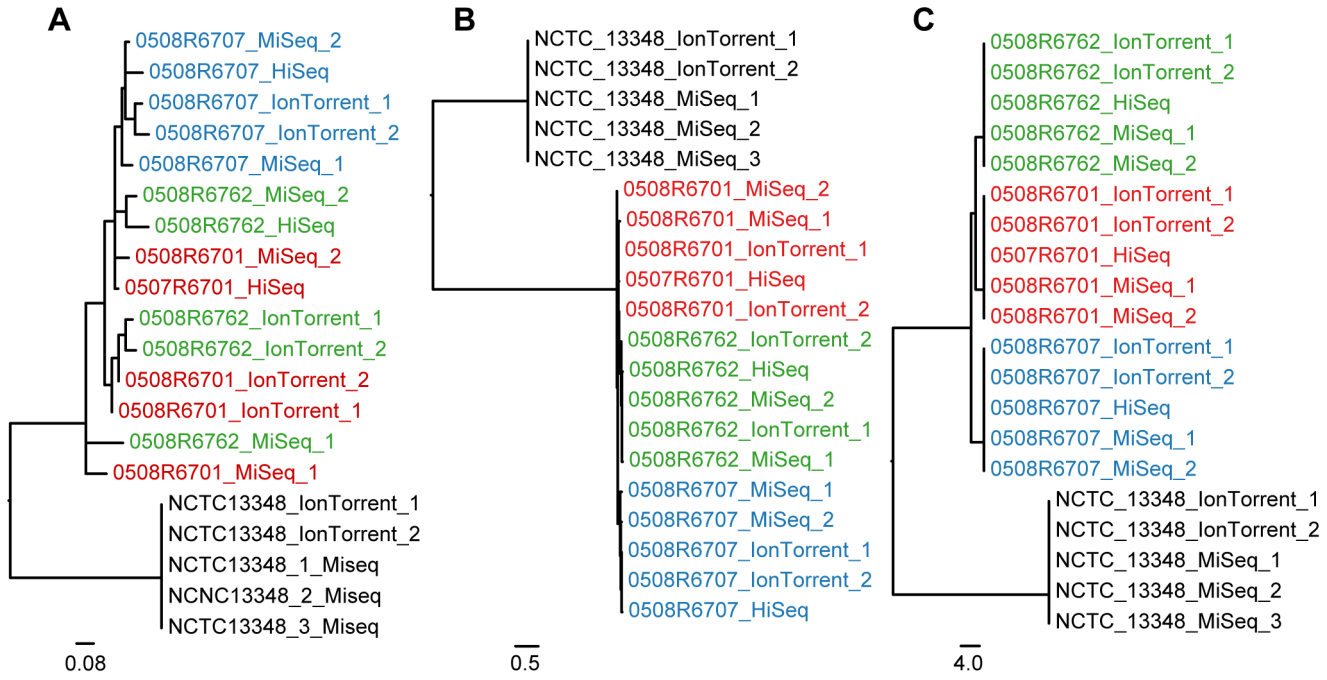

Figure 3. ***Salmonella*** **DT104 phylogeny.** Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. If repetitive sequencing has been performed then the label has also been appended either "1" or "2". (**A**) Phylogeny inferred with snpTree; (**B**) Phylogeny inferred with the novel SNP procedure; (**C**) Phylogeny inferred with the Nucleotide Difference (ND) method.
doi:10.1371/journal.pone.0104984.g003

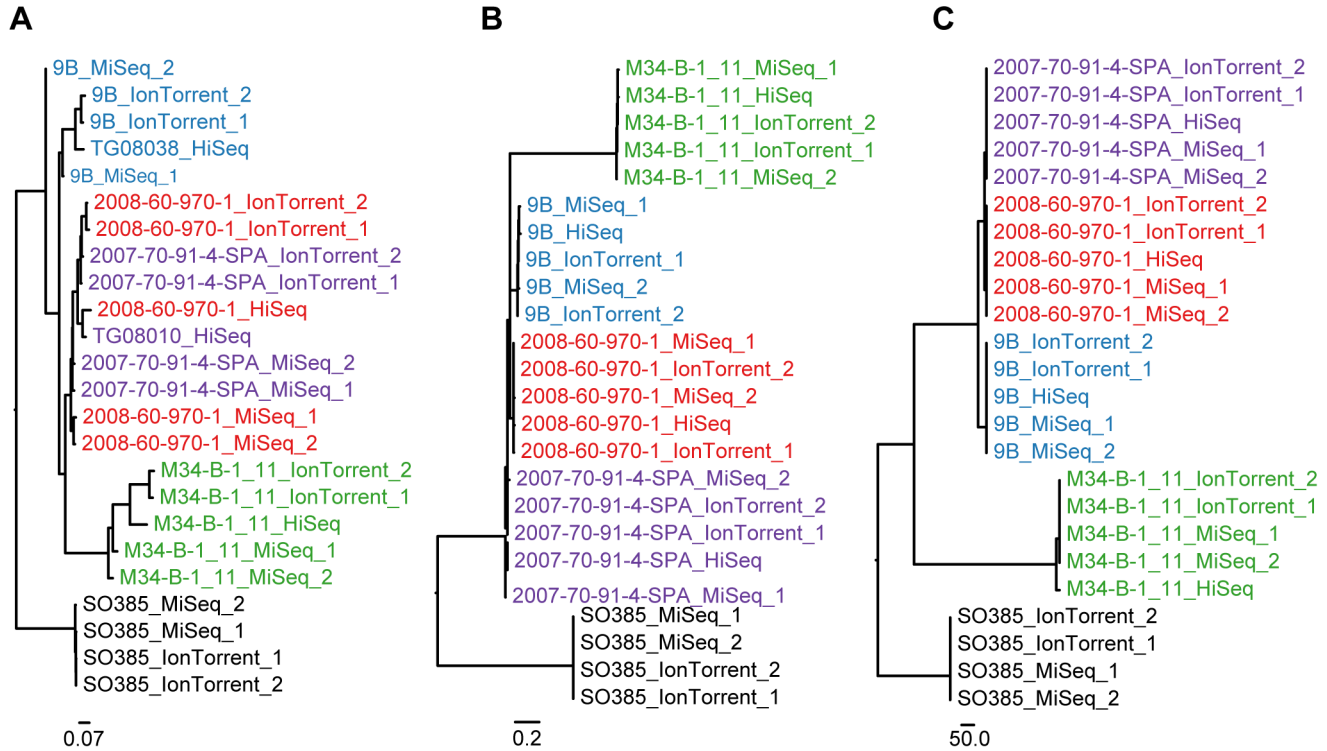

Figure 4. ***Staphylococcus aureus*** **phylogeny.** Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. If repetitive sequencing has been performed then the label has also been appended either "1" or "2". (**A**) Phylogeny inferred with snpTree; (**B**) Phylogeny inferred with the novel SNP procedure; (**C**) Phylogeny inferred with the Nucleotide Difference (ND) method.
doi:10.1371/journal.pone.0104984.g004

Figure 2A and 2B presents the phylogeny that was inferred by leaving out the three isolates with low quality sequence data. The topology generally remains the same but much more resolution is provided in these phylogenies. The increased resolution is explained by the increase of informative sites, which are doubled with the novel SNP procedure and also close to doubled with the ND method.

Salmonella Typhimurium DT104

snpTree seems to have problems differentiating properly between the sequence of the isolates that are closely related (Figure 3A), even with a closely related reference. Applying a distantly related reference a clear clustering of platforms and not isolates is seen (Figure S5). The ND method and the novel SNP procedure both cluster the isolates correctly (Figures 3C and 3B). The two methods create two identical phylogenies regardless of the distance to the reference used (see Figures S6 and S7 for phylogenies inferred with a distant reference). The novel SNP method finds between 1 and 1.5 SNPs on average between identical isolates. The ND method finds none.

Staphylococcus aureus CC398

Even with a close reference snpTree is not able to cluster the isolates 2008-60-970-1 and 2007-70-91-4-SPA correctly. These two isolates are clearly clustered according to sequencing platform and not their true relationship (Figure 4A), this clustering into sequencing platform is very clear if the distant reference is applied (Figure S8). The ND method and the novel SNP procedure both cluster the isolates correctly (Figure 4B and 4C). The ND method again infers phylogenies that are identical regardless of the distance to the reference. The novel SNP procedure infers phylogenies that are almost identical. The difference is with regard to the exact location of the node that leads to the M34-B-1_11 cluster. It is interesting that the phylogenies inferred with close references are so identical to the ones inferred by the distant references, even though the amount of informative sites increases so dramatically (see Table 2). Phylogenies inferred with a distant reference are presented in Figures S9 and S10.

Discussion

Infectious disease outbreaks often involve isolation of the causative agent in multiple laboratories within a country or even from multiple countries. Early detection of out-breaks thus, often requires rapid comparison of data from different laboratories. Next-generation sequencing shows great promises to improve the routine characterization of infectious disease agents in microbial laboratories and sequencing data are attractive because they both provide high resolution as well as a standardized data format (the DNA sequence) that may be exchanged and compared between laboratories and over time. A number of different sequencing technologies are however, available and more are expected to become available in the future. Thus, the problem with systematic biases in SNP calling between platforms may be a problem especially when, as often the cause in outbreak detection, it is necessary to identify clusters within highly similar strains.

To our knowledge we have provided the first evaluation of phylogenetic analysis done on bacterial isolates sequenced more than once and across platforms. The main reason for the success of the presented methods is in the validation of all the sites, which are part of the phylogenetic analysis. If a position is informative then that position must be called with confidence in all strains, which are part of the analysis. This validation will be very sensitive to low quality sequences. A single low quality sequencing run can cause a lot of informative sites to be ignored. However this would not cause wrong phylogenies but most likely low resolution phylogenies and the analysis, will as presented in this study clearly show which sequences to rerun or leave out and another phylogenetic analysis can quickly be done without the low quality sequences, since the mapping of read data to the reference and most of the calculations has already been done.

The presented procedures may not be perfect in identifying all single SNPs and variable sites, but for routine epidemiological typing of infectious disease agents this is less important than the correct clustering. Further evaluation also under real-time situations as done by Joensen et al. [24] are warranted, but if validated the current or modified procedures may greatly enhance our ability to compare data produced using different sequencing technologies and also provide further comparability with future technologies. The same or similar procedures might also be useful for future large-scale phylogenetic studies on human and other eukaryotic genomes.

Supporting Information

Figure S1 Ignored genome positions in novel SNP procedure (*Salmonella* Montevideo dataset). Each cluster of three columns represents the amount of genome locations that are ignored due to the addition of the specific data. Black represents MiSeq data, grey represents Ion Torrent data, and light grey represents 454 data.
(PDF)

Figure S2 *Salmonella* Montevideo phylogeny inferred by snpTree (MiSeq data only). The colors of the labels in the figure correspond to the colors used in the main figures.
(PDF)

Figure S3 *Salmonella* Montevideo phylogeny inferred by the novel SNP procedure (MiSeq data only). The colors of the labels in the figure correspond to the colors used in the main figures.
(PDF)

Figure S4 *Salmonella* Montevideo phylogeny inferred by the Nucleotide Difference method (MiSeq data only). The colors of the labels in the figure correspond to the colors used in the main figures.
(PDF)

Figure S5 *Salmonella* DT104 phylogeny inferred with snpTree (distant reference). Colors have been omitted from this figure. The sequencing platforms applied are appended to the end of each label. If repetitive sequencing has been performed then the label has also been appended either "1" or "2".
(PDF)

Figure S6 *Salmonella* DT104 phylogeny inferred with the novel SNP procedure (distant reference). Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. If repetitive sequencing has been performed then the label has also been appended either "1" or "2".
(PDF)

Figure S7 *Salmonella* DT104 phylogeny inferred with the Nucleotide Difference method (distant reference). Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. If repetitive

sequencing has been performed then the label has also been appended either "1" or "2".
(PDF)

Figure S8 *Staphylococcus aureus* phylogeny inferred with snpTree (distant reference). Colors have been omitted from this figure. The sequencing platforms applied are appended to the end of each label. If repetitive sequencing has been performed then the label has also been appended either "1" or "2".
(PDF)

Figure S9 *Staphylococcus aureus* phylogeny inferred with the novel SNP procedure (distant reference). Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. If repetitive sequencing has been performed then the label has also been appended either "1" or "2".
(PDF)

Figure S10 *Staphylococcus aureus* phylogeny inferred with the Nucleotide Difference method (distant reference). Labels are colored according to isolate. The sequencing platforms applied are appended to the end of each label. If repetitive sequencing has been performed then the label has also been appended either "1" or "2".
(PDF)

Table S1 Dataset overview.
(XLSX)

Author Contributions

Conceived and designed the experiments: RSK FMA OL PL. Performed the experiments: RSK FMA OL PL. Analyzed the data: RSK FMA OL PL. Contributed reagents/materials/analysis tools: RSK FMA OL PL. Contributed to the writing of the manuscript: RSK FMA OL.

References

1. Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, et al. (2012) Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerg Infect Dis 18: e1. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3559169&tool=pmcentrez&rendertype=abstract. Accessed 27 February 2014.
2. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13: 601–612. Available: http://www.ncbi.nlm.nih.gov/pubmed/22868263. Accessed 20 February 2014.
3. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, et al. (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 8: e1002824. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3410874&tool=pmcentrez&rendertype=abstract. Accessed 24 February 2014.
4. Harris SR, Cartwright EJP, Török ME, Holden MTG, Brown NM, et al. (2013) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis 13: 130–136. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3556525&tool=pmcentrez&rendertype=abstract. Accessed 27 February 2014.
5. Price L, Stegger M, Hasman H, Aziz M, Larsen J (2012) Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock. MBio 3: 1–6. Available: http://mbio.asm.org/content/3/1/e00305-11.short. Accessed 25 May 2012.
6. Young B, Golubchik T (2012) Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc …. Available: http://www.pnas.org/content/109/12/4550.short. Accessed 27 February 2014.
7. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, et al. (2011) Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio 2: 1–6. Available: http://mbio.asm.org/content/2/4/e00157-11.short. Accessed 27 February 2014.
8. Mellmann A, Harmsen D, Cummings C, Zentz EB, Leopold SR, et al. (2011) Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6: e22751. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3140518&tool=pmcentrez&rendertype=abstract. Accessed 2014 February 27.
9. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, et al. (2013) Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13: 137–146. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3556524&tool=pmcentrez&rendertype=abstract. Accessed 2014 February 22.
10. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, et al. (2013) Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. J Antimicrob Chemother 68: 771–777. Available: http://www.ncbi.nlm.nih.gov/pubmed/23233485. Accessed 2014 February 25.
11. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10: R32. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2691003&tool=pmcentrez&rendertype=abstract. Accessed 2014 February 20.
12. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, et al. (2012) Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30: 78–82. Available: http://www.ncbi.nlm.nih.gov/pubmed/22178993. Accessed 2014 February 27.
13. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, et al. (2013) Comparison of sequencing platforms for single nucleotide variant calls in a human sample. PLoS One 8: e55089. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3566181&tool=pmcentrez&rendertype=abstract. Accessed 2014 February 19.
14. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, et al. (2013) Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One 8: e66621. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679043&tool=pmcentrez&rendertype=abstract. Accessed 2014 January 23.
15. Suzuki S, Ono N, Furusawa C, Ying B-W, Yomo T (2011) Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS One 6: e19534. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3096631&tool=pmcentrez&rendertype=abstract. Accessed 2014 February 27.
16. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, et al. (2013) Updating benchtop sequencing performance comparison. Nat Biotechnol 31: 296. Available: http://www.ncbi.nlm.nih.gov/pubmed/23563422.
17. Allard MW, Luo Y, Strain E, Li C, Keys CE, et al. (2012) High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 13: 32. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3368722&tool=pmcentrez&rendertype=abstract. Accessed 2014 February 21.
18. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM (2014) Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica. PLoS One 9: e87991. Available: http://dx.plos.org/10.1371/journal.pone.0087991. Accessed 2014 February 5.
19. Leekitcharoenphon P, Kaas RS, Thomsen MCF, Friis C, Rasmussen S, et al. (2012) snpTree–a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics 13 Suppl 7: S6. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3521233&tool=pmcentrez&rendertype=abstract. Accessed 2014 January 21.
20. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract. Accessed 2014 November 7.
21. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832824&tool=pmcentrez&rendertype=abstract. Accessed 2013 December 12.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract. Accessed 2013 December 11.
23. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490. Available: http://www.pubmedcentral.nih.gov/articlerender.

fcgi?artid = 2835736&tool = pmcentrez&rendertype = abstract. Accessed 2011 July 28.

24. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, et al. (2014) Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol 52: 1501–1510. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3993690&tool = pmcentrez&rendertype = abstract. Accessed 2014 May 27.

Methods to define *Escherichia coli* outbreak strains based on whole genome sequence data from 10 different outbreaks.

Running title: Evaluating methods to define *E. coli* outbreak strains

Rolf S. Kaas# (rkmo@food.dtu.dk), DTU Food, The Technical University of Denmark, Kgs. Lyngby, Denmark

Simon Rasmussen (simon@cbs.dtu.dk), Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, Kgs. Lyngby, Denmark

Flemming Scheutz (fsc@ssi.dk), Statens Serum Institut, Copenhagen, Denmark

Ole Lund (lund@cbs.dtu.dk), Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, Kgs. Lyngby, Denmark

Frank M. Aarestrup (fmaa@food.dtu.dk), DTU Food, The Technical University of Denmark, Kgs. Lyngby, Denmark

Introduction

Whole genome sequencing (WGS) has become a realistic tool for real-time characterization of bacterial outbreaks and diagnostic microbiology, and is likely to become the standard procedure in the future [1]. WGS has in a number of recent studies been successfully applied for describing outbreaks and transmission events for several bacterial species including *Escherichia coli* (*E. coli*) [2,3]. In these retrospectively investigated cases there were available epidemiological data but in order for the technology to be used more routinely there is a need to establish criteria for when closely related strains should be considered clonally related.

Escherichia coli can cause large and severe outbreaks [3,4]. The outbreaks cannot always be prevented, but their impact can be significantly reduced by rapid detection of the source(s). Sequencing technology has reached a level that allows for integration of the technology in real-time surveillance [5]. However, without incorporation of WGS at the front line in the offices of the medical doctors and epidemiologists, WGS will continue to remain a purely reflective research tool. It is essential that WGS, along with the proper bioinformatics tools are made available to the people working in direct contact with patients.

If tools are to be developed that can alert the authorities to possible emerging outbreaks, then it is essential to distinguish as early as possible an "outbreak strain" from a "non-outbreak strain". Traditionally pathogenic *E. coli* is typed based on one or a subset of the following: serotype, Multi Locus Sequence Type (MLST), phylotype and Pulse Field Gel Electrophoresis (PFGE). Except for PFGE, all of these typing techniques are easily applied *in silico* from WGS, although presently only

proven for MLST [6]. Unfortunately only PFGE provides significant discriminatory power to distinguish closely related strains. Furthermore, none of these methods provides much phylogenetic information; hence they are poor candidates upon which to form the basis of automated outbreak detection.

Phylogenies inferred from Single Nucleotide Polymorphism (SNP) have in numerous studies proven to be an effective tool to classify outbreaks of several different bacterial organisms, including *E. coli* [4,7,8]. However, a number of important issues remain to be solved to make it feasible for SNP analysis to be applied in routine investigations of outbreaks. The primary concern is that the analysis depends on a reference sequence, and only sequence that can be aligned to the reference will be part of the analysis. The analysis is therefore dependent on a high quality reference sequence. For most important pathogens, these high quality reference sequences do exist, but the references for SNP analysis need to be standardized, and different versions of the references needs to be tracked (versioned) in order to provide comparable results across different analysis. Even with good standardized reference sequences, different references may yield different SNP calls and therefore potentially different results. Most studies are therefore applying reference sequences that are very closely related to the isolates being analyzed.

Another issue is SNP filtering. In most studies, SNPs are filtered out with different methods, and using different parameter settings. The filtering is in most cases done to filter out SNPs caused by sequencing errors and SNPs found in mobile genetic elements. This can be very useful but this process also needs to be standardized.

SNP analysis has become the method of choice in outbreak studies and while the method is useful, there is a lack of alternatives and a lack of benchmark testing to

show if SNP analysis really is the best method. Researchers need to also focus on the ability of the method to be applied in large scale and high throughput environments. SNP analysis has the limitation that most of the implementations applied today will require an almost complete re-calculation if new strains are added to the analysis.

We showed in a previous study that concatenated alignment of the core genes could also infer the expected phylogenies with significant resolution [9]. The method requires no reference genome, although it does require a core genome, which would also need to be standardized and versioned.

At present time, almost every WGS outbreak study is published with a different SNP method. Studies are needed that evaluate the SNP method, in an environment where the method is static, but applied to a variety of outbreaks and also measure the effect of the reference and the consequences of choosing different references for the same outbreaks. Studies are needed to benchmark alternative methods against the SNP method in order to further develop methods that will be feasible for large scale typing.

This study was conducted to evaluate different bioinformatics methodologies for analysis of WGS data and classify isolates as either outbreak or non-outbreak strains based on whole genome typing, using *Escherichia coli* as a case study. Using the sequences from 10 *E. coli* outbreaks in comparison to background strains we conducted SNP analysis, core gene analysis, average nucleotide identity (ANI) analysis, nucleotide difference (ND) analysis, and implemented a new method based on k-mers (words of size k).

Materials & Methods

Statens Serum Institut (SSI) is responsible for surveillance of *Escerichia coli* outbreaks in Denmark and provided the majority of the isolates used in this study. DTU Food provided the remaining isolates. The isolates were sequenced on the Illumina GAIIx genome analyzer or MiSeq.

A total of 46 strains were sequenced for this study; 25 of the strains were from seven different *Escherichia coli* outbreaks (See Table 1). Three of the 46 strains were not actual outbreak strains but sampled from the same individual over 3 years (referred to as the "personal" group/cluster in this study). It should be noted that the individual in question was not living in one location for the three-year period but lived in several countries. Hence the three samples have been taken in Tanzania, Egypt, and Syria.

For each of the expected clusters/outbreaks, additional strains were chosen which were unrelated but similar to the outbreak strains, these are in this study referred to as the "sporadic" or "non-outbreak" strains.

SSI has by using: epidemiological data, serotyping and PFGE defined all the outbreaks published with this study, except for the Edema disease outbreak defined by DTU Food.

It is important to note that the sequence obtained from the strain c75-10 in the Salad outbreak is of low quality. The isolate has not been removed from the analysis because it is an important indicator of how robust the different methods are.

The isolate c64-12 is in this study classified as part of the "Salad" outbreak, however the isolate was found 2 years after the actual outbreak (2012), but suspected by SSI to be clonal.

Table 1. Outbreak data. *[4]. **[3]

Outbreak	Isolates	Location	Date	Source	Serotype	MLST	Phylotype	Other
Edema	5	Denmark	1994	Pig	O139	ST-1	D	
Borupgaard	4	Denmark	2006	Human	O92	ST-1564	A	ETEC
Salad	5	Denmark	2010	Human	O6:K15:H16	ST-4	A	ETEC
Personal	3	Travel	1997-2010	Human	O117:K1:H7	ST-504	B2	
Father+Son	2	Denmark	2001	Human	O146:H21	ST-442	B1	VTEC
O157	2	Denmark	2004	Human	O157	ST-11	E	VTEC
Sandwich	4	Denmark	2012	Human	O169:H41	ST-182	D	
O104**	13	Germany, France	2011	Human	O104:H4	ST-678	B1	EAHEC
O157 taco*	2	USA	2006	Human	O157:H7	ST-11	E	VTEC
O157 spinach*	16	USA	2006	Human	O157:H7	ST-11	E	VTEC

Apart from the strains published with this study, strains from two other studies have also been included. 13 strains (+2 sporadic/historic) were included from the French O104:H4 outbreak that was connected to a larger German outbreak in 2011 [3]. 18 strains were included from O157:H7 outbreaks occurring in the United States in 2006. The American O157:H7 strains have in this study been defined as two separate outbreaks named: “O157 taco” and “O157 spinach”.

The raw data was trimmed and cleaned for adapters using AdapterRemoval v. 1.1 (https://code.google.com/p/adapterremoval/).

Two of the methods (SNP and ND) require reference genomes. The reference genomes were chosen to be either one of the outbreak strains or another closely related strain (See Table S1). All references were draft assemblies except the reference used for the outbreaks with serotype O157, where the complete sequence of the strain Sakai was employed (acc. nr. NC 002695.1).

De novo assembly (draft assembly)

VelvetK (http://bioinformatics.net.au/software.velvetk.shtml) was applied to each set of cleaned and trimmed data to estimate the k parameter for the following Velvet assembly, which gives a k-mer coverage closest to 20X.

VelvetOptimiser v. 2.2.5 (http://bioinformatics.net.au/software.velvetoptimiser.shtml) was used to test a range of k-parameters for each isolate and choose the optimal assembly. The range of k was set to the previously estimated k value +20 and -20. With a maximum of k=99 and a minimum of k=15.

Velvet v. 1.2.07 [10] was used by VelvetOptimiser to do the actual *de novo* assemblies.

SNP analysis

Trimmed and cleaned reads were mapped to reference sequences using BWA 0.5.9 [11]. The depth at each mapped position was calculated using genomeCoverageBed, which is part of BEDTools v. 2.16.2 [12]. Single Nucleotide Positions (SNPs) was called using mpileup part of SAMTools v. 0.1.18 [13]. SNPs were filtered out if the depth at the SNP position was not at least 10X or at least 10% of the average depth for the particular genome mapping. SNPs were filtered out if the mapping quality was below 25 or the SNP quality was below 30. In each mapping, SNPs are filtered out if they are called within the vicinity of 10 bp. of another SNP (pruning). A Z-score was calculated for each SNP as follows:

$$Z = (x-y)/\mathrm{sqrt}(x+y)$$

Where x is the number of reads supporting the SNP in question and y is the number of reads supporting alternate base calls or the reference base. SNPs with a Z-score < 1.96 (corresponding to a p-value of 0.05) was filtered out.

All genome mappings were then compared and all positions where SNPs was called in at least one mapping were validated in all mappings. The validation includes both the depth check and the z-score check as for the SNP filtering. Any position that fails validation is ignored in all mappings.

Raw data from the American O157 study [4] was not used, the published contigs was applied instead. Nucmer was used to align the contigs to a reference and call SNPs. The "show-snps" (with options "-CIlrT") application was used to retrieve the SNPs. Both of these applications are part of the software package MUMmer v. 3.23 [14]. The SNPs found was filtered using the previously described filters where applicable.

Maximum Likelihood trees were created using FastTree [15].

It should be noted that the genetic distance (SNP count) between a pair of isolates in the SNP analysis is different than for the distance used in the SNP cluster analysis. In the cluster analysis positions found not to be valid in all isolates are ignored, in the pair wise distance analysis only positions found not to be valid in the pair of isolates are ignored.

The method has been published [16], implemented as a web server, and is available from:

http://cge.cbs.dtu.dk/services/CSIPhylogeny/.

Core gene analysis

Prodigal v. 2.60 [17] was applied to each *de novo* assembly for gene prediction. A set of "soft-core genes" was retrieved from a previous study [9]. The soft-core genes was BLASTed [18] against the predicted genes of each genome. The genes found in all genomes were then aligned using MUSCLE v. 3.8.31 [19] and concatenated.

A BLAST hit was considered valid if the identity of the hit was at least 98% and the length of the alignment between the hit and the database gene was covering at least 98%. Thresholds of 80% and 50% were also tested, but produced inferior trees compared to the final thresholds used (data not shown).

DNADist (part of the PHYLIP package [20]) was used to calculate the genetic distances from the multiple alignments and FastMe [21] was used to calculate the final trees from the distance matrices.

K-mer analysis

Each isolate was assembled *de novo*. All possible k-mers of length 35 were found from the assemblies. The number of 35-mers shared between each pair of isolates was counted. 35-mers matching several positions in an isolate was only counted once. For each pair the genetic distance (kmer_dist) was calculated as:

$$kmer\_dist = 1 - ((s / t1) + (s / t2)) / 2$$

Where "s" is the number of shared 35-mers between the two isolates and t1 and t2 are the total number of different 35-mers found in each of the two isolates. The distance thus represents the average percentage of 35-mers that are different between the two isolates. Trees where calculated from the genetic distances using FastME [21].

K-mers between the lengths of 5 and 500 was considered (See Figure S1 before it was decided to use the length of 35.

Nucleotide Difference (ND) analysis

The reference genome was split into k-mers of length 17 and stored in a hash table. Each read with a length of at least 50 was split in to 17-mers overlapping by 16. K-mers from the read and its reverse complement was mapped until an ungapped

alignment with a score of at least 50 was found using a match score of 1 and a mismatch score of -3.

When all reads had been mapped, the significance of the base call at each position was evaluates by calculating the number of reads x having the most common nucleotide at that position, and the number of reads y supporting other nucleotides. A Z-score was calculated as:

$$Z = (x-y)/sqrt(x+y)$$

The value of 3.29 was used as a threshold for Z corresponding to a p-value of 0.001. It was further required that x>10*y.

Each pair of sequences was compared and the number of nucleotide differences at all positions called in all of the strains to be compared was counted. A matrix with these numbers was given as input to an UPGMA algorithm implemented in the neighbor program (part of the PHYLIP package [20]) in order to construct the tree.

The method has been published [22],implemented as a web server, and is available at: http://cge.cbs.dtu.dk/services/NDtree/.

Average Nucleotide Identity (ANI)

This method was suggested as an *in silico* method for DNA-DNA hybridization by Goris et al. [23].The method was implemented as described in the study. The method provides the percentage of similar DNA between two isolates. The results were reversed in order to get a percentage of dissimilarity. These percentages were then used as genetic distances and trees were created using FastME [21].

Results

Clustering

Phylogenies were inferred using the 5 different methods described in the section Materials and Methods. Each expected outbreak cluster was investigated. A cluster was considered correct if the outbreak strains formed a monophyletic clade (only contained the outbreak strains and no other strains). All methods were able to cluster all the outbreaks correctly except for two of the outbreaks (see Table 2).

The SNP method was able to cluster all the outbreaks correctly assuming the reference strain used was closely related (closely related reference sequences for each outbreak is specified in Table 2). If a distantly related reference strain were used, then the SNP method would still cluster most outbreaks correctly, except one, the "Edema" outbreak. The ANI method failed to produce correct clusters for the "Edema" and the "O157 spinach" outbreak.

Table 2. Failed clustering. "+" equals successful clustering and "-" equals failed clustering. Parentheses means only successful if a close reference was employed. Clustering results not indicated in this table were all successful.

	Outbreak	
Method	**Edema**	**O157 spinach**
Single Nucleotide Polymorphism (SNP)	(+)	+
K-mer	+	+
Nucleotide Difference (ND)	+	+
Core genes	+	+
Average nucleotide identity (ANI)	-	-

Genetic distance between outbreak and non-outbreak strains

To assess the ability of the methods to differentiate between outbreak strains and non-outbreak strains, the genetic distances computed by the methods between the strains are measured. For each group of outbreak isolates, the average distance between the strains within an outbreak is calculated. The maximum distance between any pair of outbreak strains are measured and the minimum distance from any outbreak strain to a

non-outbreak strain is found. The results for each of the methods are shown in Figure 1. The specific numbers that make up Figure 1 can be found in Table S2.

Only the SNP method produced genetic distances that made it possible to define a static threshold differentiating outbreak strains from non-outbreak strains in the dataset applied for this study (the area of possible thresholds are marked with a red horizontal bar in Figure 1). The limit of the lowest possible threshold is 496 SNPs and its defined by the Edema outbreak. The personal cluster will define the lower limit if only the averages are considered (461 SNPs). The limit of the upper threshold is 601 SNPs and is defined by the American O157 outbreaks.

The results of the ND method are very similar to the ones from the SNP method, except for the two American O157 outbreaks that comprises only assembled genomes. Ignoring the American O157 outbreaks the range of thresholds would be 771-1065 nucleotide differences.

The low quality “Salad” strain c75-10 is causing a large effect in the three methods: k-mer, core, and ANI. The maximum distance within the Salad outbreak is unexpectedly high, and is in all three methods caused by the c75-10 strain.

The k-mer method generally finds larger variation within outbreaks than the other methods. A static threshold cannot be defined. Investigating each outbreak and the corresponding sporadic cases (nearest neighbors) shows that an individual threshold for each outbreak can be defined – a dynamic threshold, with the exception of the O157 spinach outbreak, which has a single isolate that causes a threshold to fail. This is also true for the ND method.

Figure 1. Genetic distance between outbreak and background. Green bars indicate variance found within each outbreak (dark=average, light=max). Blue bars indicate distance to nearest non-outbreak strain. Each blue bar reaching the top expands beyond view. The red bar (horizontal) indicates the clonal threshold for the SNP analysis.

The core gene method generally finds less variation between outbreak strains and non-outbreak strains. A dynamic threshold can be defined for this method, also for the O157 spinach outbreak, which caused difficulties for the k-mer method.

Neither static nor a dynamic threshold can be set for the ANI method. This is not surprising due to the difficulties also experienced in the clustering analysis.

Discussion

In this study 5 different bioinformatics methods were applied to WGS of 77 strains including 10 different outbreaks/clusters, of which 25 strains from 7 outbreaks and 21 sporadic strains were sequenced for this study. The genetic distances obtained from the different methods were measured with a focus on differentiating outbreak strains from non-outbreak strains in the future enabling this technology to be applied in further development of automated outbreak detection using WGS data.

Except for the Personal strains, all outbreaks presented here lasted only a few months. No long lasting *E. coli* outbreaks was included because the majority of *E. coli* outbreaks, even though they can be quite large and have severe implications are relatively short.

The current study does not have focus on phylogeny but the ability of the different methods to cluster isolates correctly. All the methods applied in this study were able to cluster most of the outbreaks correctly. Only two outbreaks were not clustered correctly by one of the methods, ANI (see Table 2). The SNP method did also fail to cluster the Edema outbreak correctly but only if a distant reference was applied.

In general, all methods except ANI managed to make a clear distinction between the different outbreaks and the sporadic/background strains in the study, with the exception of the O157 spinach outbreak (see Figure 1). O157:H7 does have (compared to other serotypes) an unusual homogenous population structure [4] and might have to be considered a special case, in clustering analysis.

The most widely used method for analyzing outbreaks is SNP analysis and it has been applied in several studies including E. coli [2,3,5]. In previous studies a within outbreak variation of up to 74 SNPs has been observed [4]. In this study the SNP analysis suggest that *E. coli* strains with less than 500 SNP differences may be related to the same outbreak (Figure 1). The static threshold presented in Figure 1 is very specific for this particular dataset. It could be argued that the Personal cluster should be left out because it runs over 3 years and therefore is expected to contain more diversity than the other outbreaks that only runs over a couple of months. The Edema disease outbreak was also of longer duration and if these two clusters are taken out the threshold becomes 200 SNPs. The implication of leaving out these two outbreaks is even higher on the ND method (if the American O157 outbreaks are ignored), where the threshold drops from around 800 down to about 200 differences. Both the SNP and the ND method rely on a reference sequence. It is widely accepted that a close reference sequence is needed for SNP analysis, which is also confirmed by this study. It is hypothesized that using a reference to calculate genetic distances is likely to be the reason why these methods provide a larger differentiation between outbreak strains and non-outbreak strains. The number of SNPs or nucleotide differences found in an isolate using a distant reference is very large and each SNP or nucleotide difference is no longer representing a single evolutionary event. A phylogeny inferred

between distantly related strains is therefore less reliable, but this effect could explain why these methods are able to create more significant differentiations, even though another method might provide more correct genetic distances.

It has previously been described by Touchon et al. [24] that even though *E. coli* has a very dynamic genome, it is only dynamic in certain hotspots and seems to retain a stable core, which gives it a clonal nature. It explains the rather low genetic distances found, using the core method, between outbreak strains and non-outbreak strains. From the results of the Salad outbreak the core method also seems less robust against poor sequencing. However, the robustness might be increased by a more sophisticated core gene method that have more focus on eliminating sequence errors.

The k-mer method is also vulnerable to sequencing errors and in general seems to calculate higher genetic distances between outbreak strains than the other methods. It is believed that the explanation is how the k-mer method uses the entire pan-genome and not only the core genome sequence. Furthermore, a good method for sorting out sequencing errors has not yet been found for this method.

The ND method seems to have results relatively comparable to the ones from the k-mer method for the two American O157 outbreaks. These outbreaks causes much more variation within outbreaks than is seen for the outbreaks that comprises only raw read data. Less variation would be expected in the O157 outbreaks due to the lower variation seen in the SNP analysis. The ND method was initially developed to only handle raw read data, and the results suggests that further development of the assembled data part of the ND method might improve the method. However, it should be noted that the method was able to cluster the outbreaks correctly (See supplementary figure S4)

The ANI method was developed with focus on species differentiation and not inter-species differentiation, which explains why it struggles with resolution.

None of the methods applied in this study presented any major weaknesses. This confirms, as expected, that the SNP method is a very reliable method but also that feasible alternatives can be developed. The SNP method is undoubtedly the most mature of the methods presented and in this study also slightly superior to the other methods, regarding outbreak and non-outbreak differentiation. However, it is believed that both the ND method and the K-mer method can be implemented more effectively than a SNP analysis, both due to the programmatic approach but also due to the ability of these methods to apply new strains without recalculating all the genetic distances between all strains.

In future large-scale environments both speed and comparability will be extremely important and SNP analysis might not be feasible. However, combining a crude fast method with a SNP analysis might provide the solution. Alternative methods need to be developed and further benchmarking studies is needed to ensure the best methods are applied.

References

1. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med [Internet]. **2012**; 366(24):2267–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22693998
2. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, et al. Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. J Clin Microbiol [Internet]. **2013** [cited 2014 Jun 17]; 51(1):232–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3536255&tool=pmcentrez&rendertype=abstract
3. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, FitzGerald M, et al. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci [Internet]. **2012** [cited 2012 Feb 7]; Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1121491109
4. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of Escherichia coli O157: H7 outbreaks. Proc Natl Acad Sci [Internet]. National Acad Sciences; **2011** [cited 2012 Jan 5]; 108(50):20142–20147. Available from: http://www.pnas.org/content/108/50/20142.short
5. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol [Internet]. **2014** [cited 2014 May 27]; 52(5):1501–10. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3993690&tool=pmcentrez&rendertype=abstract
6. Larsen M V, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus Sequence Typing of Total Genome Sequenced Bacteria. J Clin Microbiol [Internet]. **2012** [cited 2012 Mar 17]; 50(4):1355–1361. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22238442
7. Harris SR, Feil EJ, Holden MTG, Quail M a, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science [Internet]. **2010**; 327(5964):469–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20093474
8. Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med [Internet]. **2011**; 364(1):33–42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3030187&tool=pmcentrez&rendertype=abstract
9. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genomics [Internet]. **2012** [cited 2013 Nov 13]; 13:577. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3575317&tool=pmcentrez&rendertype=abstract
10. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res [Internet]. **2008** [cited 2013 Nov 6]; 18(5):821–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2336801&tool=pmcentrez&rendertype=abstract

11. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. **2009** [cited 2013 Nov 7]; 25(14):1754–60. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract
12. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics [Internet]. **2010** [cited 2013 Dec 12]; 26(6):841–2. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832824&tool=pmcentrez&rendertype=abstract
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. **2009** [cited 2013 Dec 11]; 25(16):2078–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract
14. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res [Internet]. **2002**; 30(11):2478–83. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=117189&tool=pmcentrez&rendertype=abstract
15. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One [Internet]. **2010** [cited 2011 Jul 28]; 5(3):e9490. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2835736&tool=pmcentrez&rendertype=abstract
16. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. Friedrich A, editor. PLoS One [Internet]. **2014** [cited 2014 Aug 11]; 9(8):e104984. Available from: http://dx.plos.org/10.1371/journal.pone.0104984
17. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics [Internet]. **2010**; 11:119. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2848648&tool=pmcentrez&rendertype=abstract
18. Altschul S, Gish W, Miller W. Basic local alignment search tool. J Mol … [Internet]. **1990** [cited 2013 Dec 12]; 215(3):403–410. Available from: http://linkinghub.elsevier.com/retrieve/doi/10.1006/jmbi.1990.9999
19. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res [Internet]. **2004** [cited 2011 Jun 10]; 32(5):1792–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=390337&tool=pmcentrez&rendertype=abstract
20. Felsenstein J. PHYLIP - Phylogeny Inference Package. Cladistics. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.; **1989**; 5:164–166.
21. Desper R, Gascuel O. Fast and Accurate Phylogeny Minimum-Evolution Principle. J Comput Biol. **2002**; 9(5):687–705.
22. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. Evaluation of Whole Genome Sequencing for Outbreak Detection of

Salmonella enterica. Chabalgoity JA, editor. PLoS One [Internet]. **2014** [cited 2014 Feb 5]; 9(2):e87991. Available from: http://dx.plos.org/10.1371/journal.pone.0087991

23. Goris J, Konstantinidis KT, Klappenbach J a, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol [Internet]. **2007** [cited 2012 Mar 14]; 57(Pt 1):81–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17220447
24. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet [Internet]. **2009** [cited 2011 Jul 21]; 5(1):e1000344. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2617782&tool=pmcentrez&rendertype=abstract

Supplementary Material

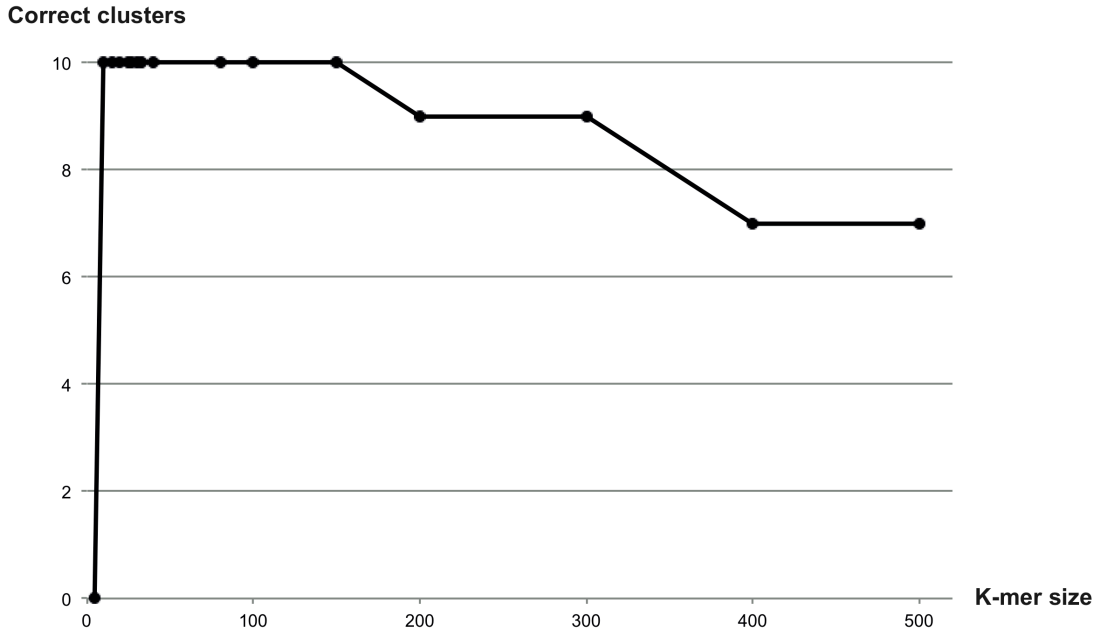

Figure S1: Evaluation of k-mer analysis with different values of k. Phylogenies was created based on k-mers for different values of k and the number of correct clusters was plotted.

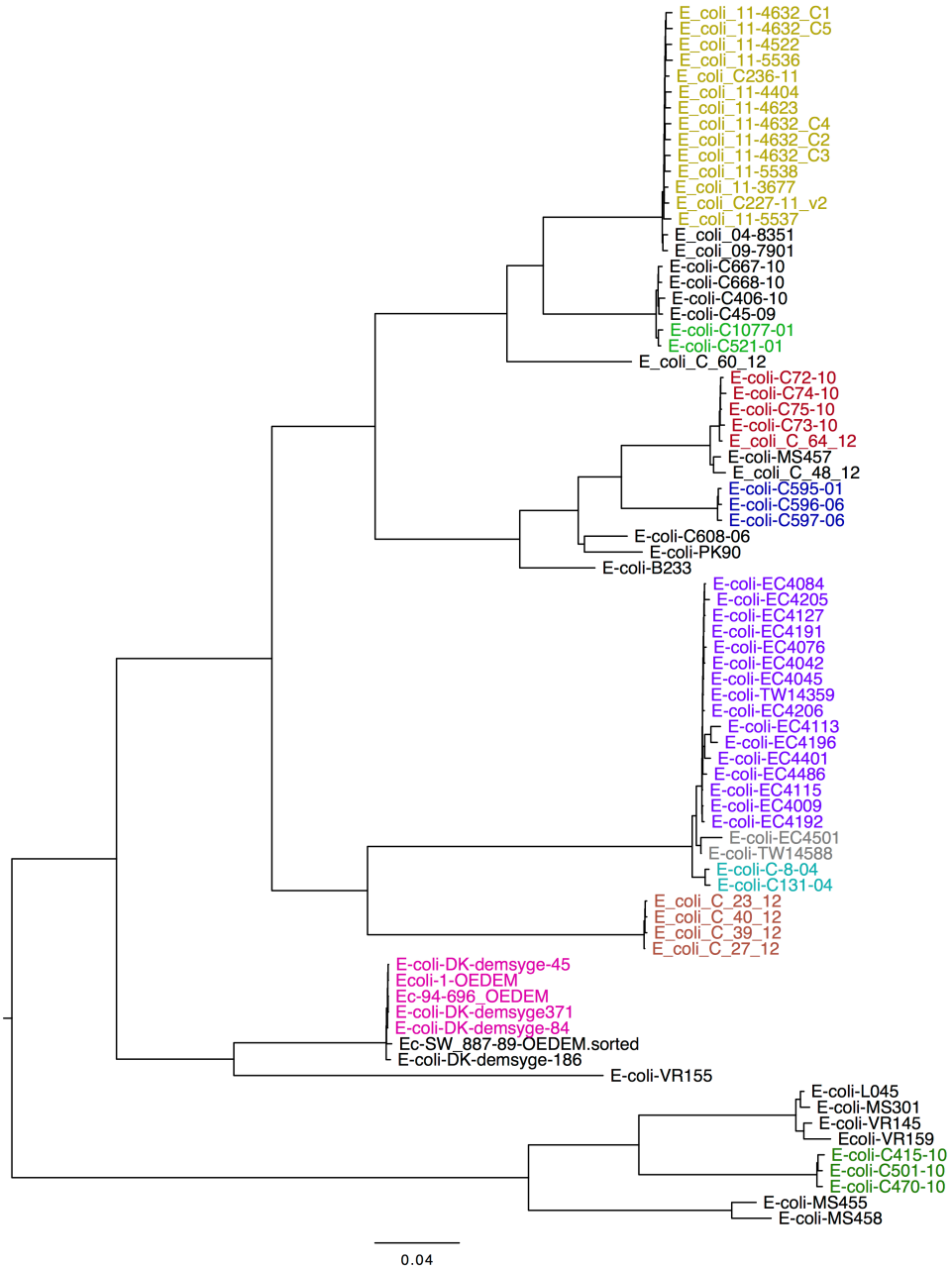

Figure S2: SNP tree. Maximum likelihood phylogeny created using reference strain "Oedem_94_dk" from the Edema outbreak. Each color represents a specific outbreak. *Yellow*: O104 outbreak. *Green*: Father+Son outbreak. *Red*: Salad outbreak. *Blue*: Borupgaard. *Purple*: O157 spinach outbreak. *Grey*: O157 taco outbreak. *Light blue*: O157. *Brown*: Sandwich outbreak. *Pink*: Edema outbreak. *Green*: Personal cluster.

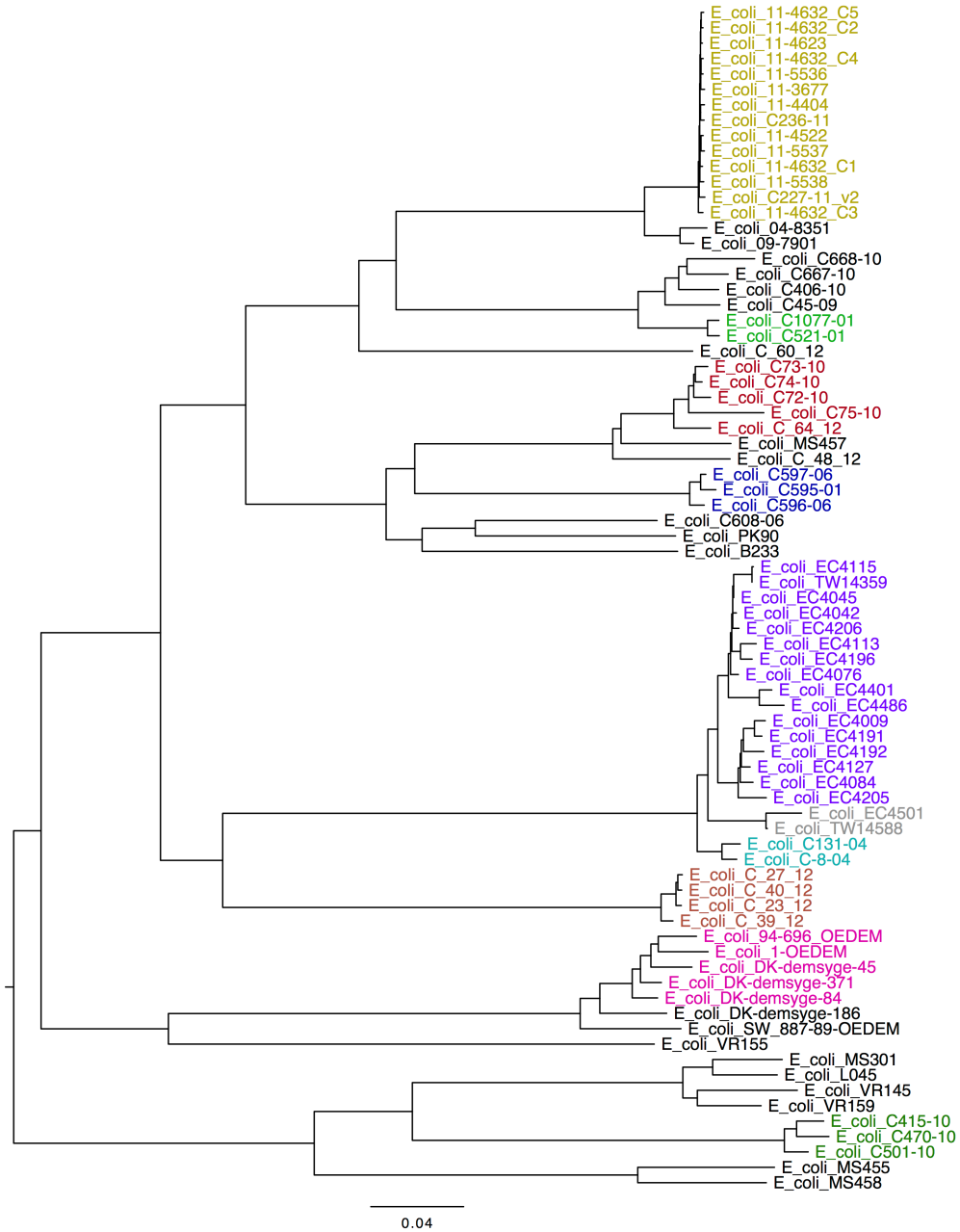

Figure S3: K-mer tree. FastME phylogeny inferred from k-mers. Each color represents a specific outbreak. *Yellow*: O104 outbreak. *Green*: Father+Son outbreak. *Red*: Salad outbreak. *Blue*: Borupgaard. *Purple*: O157 spinach outbreak. *Grey*: O157 taco outbreak. *Light blue*: O157. *Brown*: Sandwich outbreak. *Pink*: Edema outbreak. *Green*: Personal cluster.

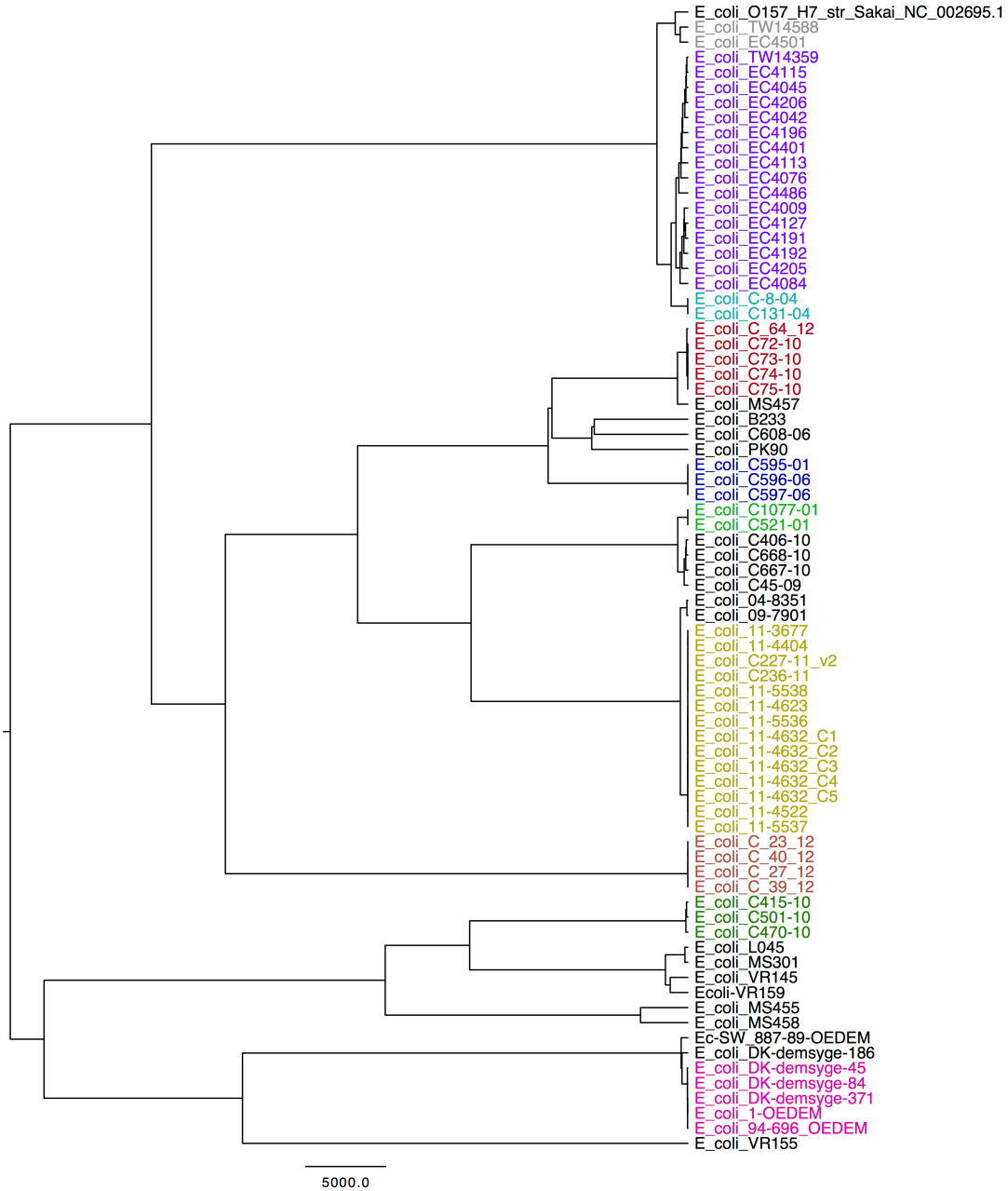

Figure S4: Nucleotide Difference (ND) tree. UPGMA phylogeny created using reference strain "TY-2482" from the O104 outbreak. Each color represents a specific outbreak. *Yellow*: O104 outbreak. *Green*: Father+Son outbreak. *Red*: Salad outbreak. *Blue*: Borupgaard. *Purple*: O157 spinach outbreak. *Grey*: O157 taco outbreak. *Light blue*: O157. *Brown*: Sandwich outbreak. *Pink*: Edema outbreak. *Green*: Personal cluster.

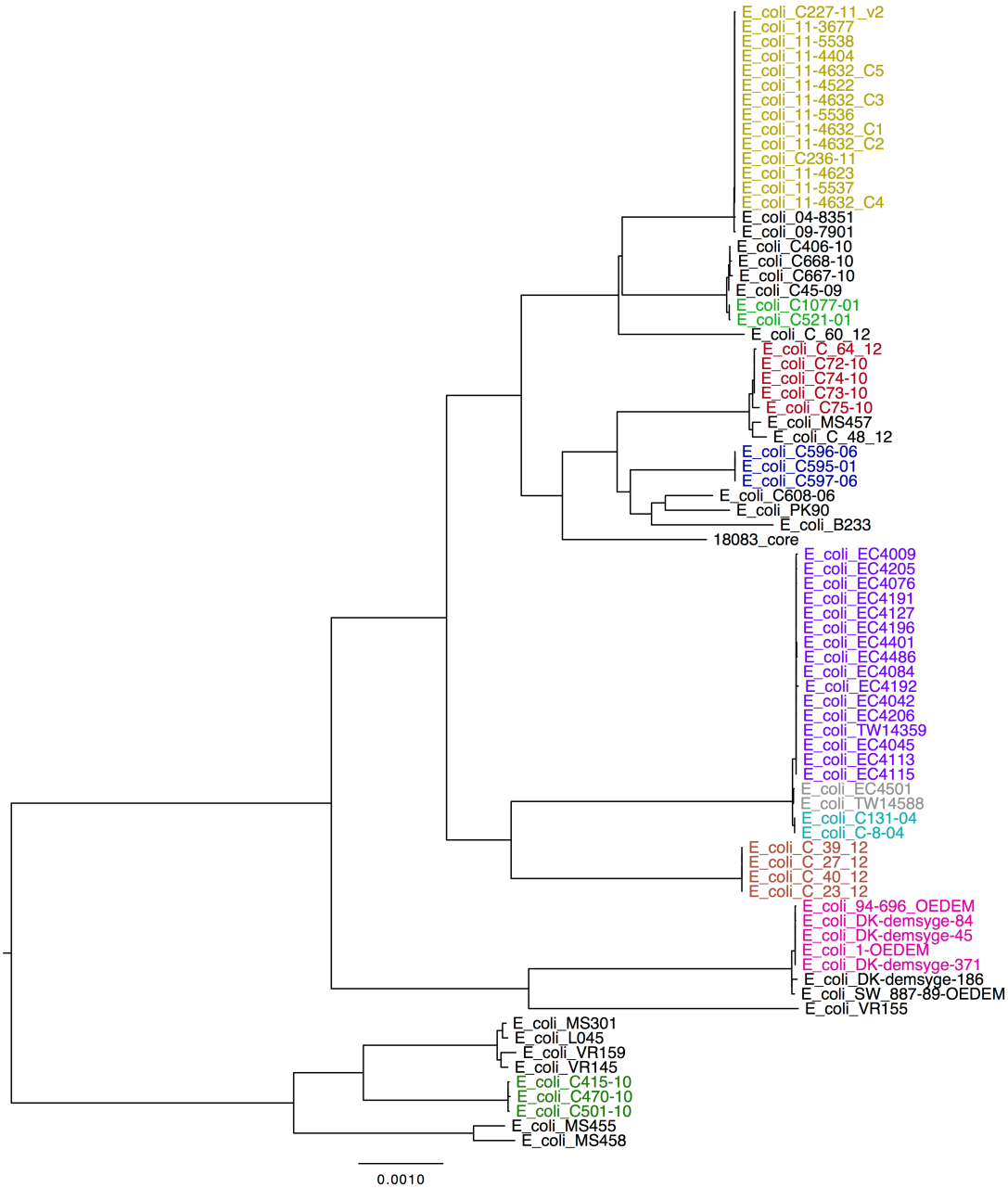

Figure S5: Core gene tree. Maximum likelihood phylogeny inferred from the soft-core genome. Each color represents a specific outbreak. *Yellow*: O104 outbreak. *Green*: Father+Son outbreak. *Red*: Salad outbreak. *Blue*: Borupgaard. *Purple*: O157 spinach outbreak. *Grey*: O157 taco outbreak. *Light blue*: O157. *Brown*: Sandwich outbreak. *Pink*: Edema outbreak. *Green*: Personal cluster.

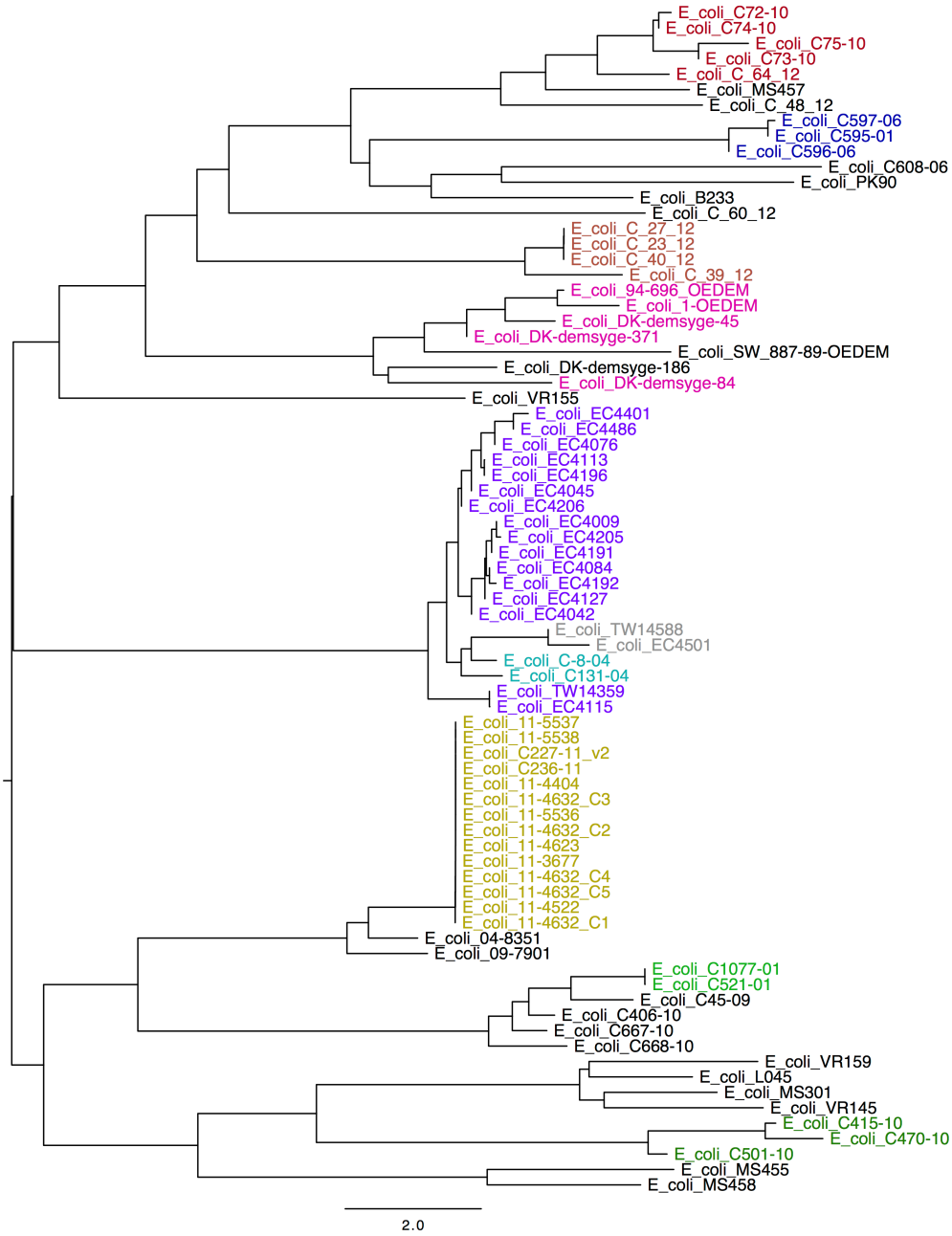

Figure S6: Average Nucleotide Identity (ANI) tree. FastME phylogeny inferred from the average nucleotide differences found between all strain pairs. Each color represents a specific outbreak. *Yellow*: O104 outbreak. *Green*: Father+Son outbreak. *Red*: Salad outbreak. *Blue*: Borupgaard. *Purple*: O157 spinach outbreak. *Grey*: O157 taco outbreak. *Light blue*: O157. *Brown*: Sandwich outbreak. *Pink*: Edema outbreak. *Green*: Personal cluster.

Table S1: Table of references used in this study.

Cluster/Outbreak	Reference strain	Reference data source
Edema	Oedem_94_dk	Current study
Borupgaard	C 608-06	Current study
Salad	C 74-10	Current study
Personal	C 501-10	Current study
Father+Son	C 521-01	Current study
O157	C 8-04	Current study
Sandwich	C 39-12	Current study
O104	TY-2482	https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/blob/master/strains/TY2482/seqProject/BGI/assemblies/BGI/Escherichia_coli_TY-2482.chromosome.20110616.fa.gz
O157 taco	O157:H7 str. Sakai	RefSeq: NC_002695.1
O157 spinach	O157:H7 str. Sakai	RefSeq: NC_002695.1

Table S2: Values for the histograms in Figure 1.

	Edema			Borupgaard			Salad		
Method	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**
SNP	215	496	1146	25	46	11674	84	203	2026
K-mer	0,0367	0,0447	0,0525	0,0132	0,0182	0,2341	0,0295	0,0565	0,0838
ND	156	771	1143	3	5	8775	66	164	1065
Core	0,E+00	0,E+00	4,E-05	5,E-05	1,E-04	2,E-04	0,E+00	0,E+00	6,E-03
ANI	2	6	2	0,5	2	8	1,35	4	3

	O157			Sandwich			O104		
Method	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**
SNP	16	16	751	2	3	30489	6	12	695
K-mer	0,0142	0,0142	0,0327	0,0093	0,0146	0,4118	0,0033	0,0053	0,0454
ND	7	7	22270	2	3	30533	5	10	1459
Core	9,E-07	6,E-06	2,E-05	5,E-06	1,E-05	7,E-05	3,E-05	4,E-05	3,E-03
ANI	0,5	1	1	0,5	2	9	0	0	2

	Personal			Father+Son		
Method	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**
SNP	380	461	19585	40	40	1562
K-mer	0,0274	0,0300	0,3204	0,0099	0,0099	0,0677
ND	374	417	24769	4	4	1883
Core	6,E-06	6,E-06	4,E-05	0,E+00	0,E+00	2,E-03
ANI	1,33	3	10	0	0	2

	O157 taco			O157 spinach		
Method	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**
SNP	92	92	601	59	119	601
K-mer	0,0159	0,0159	0,0338	0,0253	0,0556	0,0327
ND	949	949	1419	1097	1994	1339
Core	8,E-06	4,E-05	6,E-05	6,E-06	6,E-06	6,E-05
ANI	0,5	1	2	1,04	6	1

Table S3: Strain data sequenced for current study (2 pages).

Outbreak	Strain	N50	Contigs	Location
Edema	1_oedem	98148	296	Denmark
Edema	371_oedem	99476	287	Denmark
Edema	94-696_oedem	110503	273	Denmark
Edema	oedemsyge-45	89727	321	Denmark
Edema	oedemsyge-84	97011	223	Iceland
Borupgaard	C 596-06	73806	482	Denmark, Borupgaard
Borupgaard	C 597-06	75924	673	Denmark, Borupgaard
Borupgaard	C 598-06	60801	548	Denmark, Borupgaard
Borupgaard	C 608-06	147371	239	Denmark, Borupgaard
Salad	C 72-10	88562	324	Denmark
Salad	C 73-10	91929	287	Denmark
Salad	C 74-10	103771	285	Denmark
Salad	C 75-10	28149	1046	Denmark
Salad	C 64-12	111799	354	Denmark
Personal	C 415-10	59634	443	Tanzania
Personal	C 470-10	62291	363	Egypt
Personal	C 501-10	63864	315	Syria
Father+Son	C 1077-01	100587	643	Denmark
Father+Son	C 521-01	110377	632	Denmark
O157	C 131-04	133118	516	Denmark
O157	C 8-04	156872	403	Denmark
Sandwich	C 23-12	250736	126	Denmark
Sandwich	C 27-12	214064	131	Denmark
Sandwich	C 39-12	250803	94	Denmark
Sandwich	C 40-12	250601	120	Denmark
sporadic	C 60-12	109003	325	Denmark
sporadic	oedemsyge-186	96664	316	Unknown
sporadic	SW887/89 2.93 oedem	115337	221	Switzerland
sporadic	VR155	113445	386	Denmark, Hvidovre
sporadic	PK90	87000	370	Unknown
sporadic	B233	92282	320	China
sporadic	MS457	88700	549	Unknown
sporadic	C 48-12	73452	279	Denmark
sporadic	MS455	134190	229	Unknown
sporadic	MS301	175472	364	Unknown
sporadic	MS458	62062	428	Unknown
sporadic	L045	169603	305	China
sporadic	VR145	214324	360	Denmark, Hvidovre
sporadic	VR159	202859	129	Denmark, Hvidovre
sporadic	C 406-10	160258	595	Denmark
sporadic	C 45-09	177395	537	Unknown
sporadic	C 667-10	155704	451	Africa
sporadic	C 668-10	178336	526	Kenya

Source	Date	Serotype	Phylotype	MLST	Virulence
Pig	1994	O139	D	ST-1	Edema disease
Pig	1994	O139	D	ST-1	Edema disease
Pig	1994	O139	D	ST-1	Edema disease
Pig	1994	O139	D	ST-1	Edema disease
Pig	1994	O139	D	ST-1	Edema disease
Human	2006	O92:H-	A	ST-1564	ETEC
Human	2006	O92:H-	A	ST-1564	ETEC
Human	2006	O92:H-	A	ST-1564	ETEC
Human	2006	O153:H2	A	ST-10	ETEC
Human	2010	O6:K15:H16	A	ST-4	ETEC
Human	2010	O6:K15:H16	A	ST-4	ETEC
Human	2010	O6:K15:H16	A	ST-4	ETEC
Human	2010	O6:K15:H16	A	ST-4	ETEC
Human	2012	O6:K15:H16	A	ST-4	ETEC
Human	1997-2010	O117:K1:H7	B2	ST-504	VTokEPI
Human	1997-2010	O117:K1:H7	B2	ST-504	VTokEPI
Human	1997-2010	O117:K1:H7	B2	ST-504	VTokN
Human	2001	O146:H21	B1	ST-442	VTEC
Human	2001	O146:H21	B1	ST-442	VTEC
Human	2004	O157:H-	D	ST-11	VTEC
Human	2004	O157:H-	D	ST-11	VTEC
Human	2012	O169:H41	D	ST-182	ETEC
Human	2012	O169:H41	D	ST-182	ETEC
Human	2012	O169:H41	D	ST-182	ETEC
Human	2012	O169:H41	D	ST-182	ETEC
Human	2012	O169:H41	B1	ST-1490	ETEC
Pig	Unknown	O139	D	ST-1	Edema disease
Pig	1989	O139	D	ST-1	Edema disease
Human	Unknown		D*	ST-69	Pyelonephritic
Human	Unknown		A	ST-10	
Human	Unknown		A*	ST-44	
Human	Unknown		A	ST-4	ETEC
Human	2012	O6:K15:H16	A	ST-4	ETEC
Human	Unknown		B2	ST-15	ETEC
Human	Unknown		B2*	ST-73	Pyelonephritic
Calf	Unknown		B2	ST-19	EPEC
Human	Unknown		B2*	ST-73	
Human	Unknown		B2*	ST-73	Cystitis
Human	Unknown		B2*	ST-73	Cystitis
Human	2010	O146:H21	B1	ST-442	VTokEPI
Human	Unknown	O146:H21	B1	ST-442	VTokEPI
Human	Unknown	O146:H21	B1	ST-442	VTokEPI
Human	Unknown	O146:H21	B1	ST-442	VTokN

Table S4: Strain data from other studies (2 pages).

Outbreak	Strain	Location	Source	Date	Serotype
O104	11-4404	France	Human	2011	O104:H4
O104	11-4522	France	Human	2011	O104:H4
O104	11-4623	France	Human	2011	O104:H4
O104	11-4632_C1	France, Bordeaux	Human	2011	O104:H4
O104	11-4632_C2	France, Bordeaux	Human	2011	O104:H4
O104	11-4632_C3	France, Bordeaux	Human	2011	O104:H4
O104	11-4632_C4	France, Bordeaux	Human	2011	O104:H4
O104	11-4632_C5	France, Bordeaux	Human	2011	O104:H4
O104	11-5536	France, Bordeaux	Human	2011	O104:H4
O104	11-5537	France, Bordeaux	Human	2011	O104:H4
O104	11-5538	France, Bordeaux	Human	2011	O104:H4
O104	11-3677	Germany	Human	2011	O104:H4
O104	11-3798	Germany	Human	2011	O104:H4
O104	C236-11	Denmark	Human	2011	O104:H4
O104	C227-11_v2	Denmark	Human	18/05/11	O104:H4
O157 Taco	EC4501	US	Human	Nov-06	O157:H7
O157 Taco	TW14588	US	Lettuce	2006	O157:H7
O157 Spinach	EC4486	US	Human	2006	O157:H7
O157 Spinach	EC4401	US	Human	2006	O157:H7
O157 Spinach	EC4205	US	Bovine		O157:H7
O157 Spinach	TW14359	US	Human		O157:H7
O157 Spinach	EC4084	US	Human		O157:H7
O157 Spinach	EC4127	US	Human		O157:H7
O157 Spinach	EC4191	US	Spinach bag		O157:H7
O157 Spinach	EC4076	US	Human		O157:H7
O157 Spinach	EC4113	US	Spinach bag		O157:H7
O157 Spinach	EC4042	US	Human		O157:H7
O157 Spinach	EC4045	US	Spinach bag		O157:H7
O157 Spinach	EC4206	US	Bovine		O157:H7
O157 Spinach	EC4196	US	Bovine		O157:H7
O157 Spinach	EC4115	US	Human		O157:H7
O157 Spinach	EC4192	US	Human		O157:H7
O157 Spinach	EC4009	US	Human		O157:H7
sporadic	04-8351	France	Human	2004	O104:H4
sporadic	09-7901	France	Human	2009	O104:H4

Phylotype	MLST	Virulence
B1	ST-678	EAHEC
B1	ST-678	EAHEC
B1	ST-678	EAHEC
B1	ST-678	EAHEC
B1	Unknown	EAHEC
B1	Unknown	EAHEC
B1	Unknown	EAHEC
B1	ST-678	EAHEC
B1	Unknown	EAHEC
B1	Unknown	EAHEC
B1	ST-678	EAHEC
B1	ST-678	EAHEC
B1	ST-678	EAHEC
B1	Unknown	EAHEC
B1	ST-678	EAHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11	EHEC
E	ST-11*	EHEC
E	ST-11	EHEC
B1	ST-678	EAHEC
B1	ST-678	EAHEC